



Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н. Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н. Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ

НА ТЕМУ:

«Оптимизация метода сжатия страниц памяти с
использованием подсчета информационной энтропии»

Студент ИУ7-83Б
(Группа)

(Подпись, дата)

Р. Р. Хамзина
(И. О. Фамилия)

Руководитель ВКР

(Подпись, дата)

А. А. Оленев
(И. О. Фамилия)

Нормоконтролер

(Подпись, дата)

Д. Ю. Мальцева
(И. О. Фамилия)

2023 г.

РЕФЕРАТ

Расчетно-пояснительная записка 56 с., 23 рис., 4 табл., 37 источн., 1 прил.

В данной работе представлена разработка оптимизации метода сжатия страниц памяти с использованием подсчета информационной энтропии.

В разделе 1 представлено описание основных механизмов подсистемы памяти операционной системы, изучены сжатие данных и показатели его качества, рассмотрены понятие информационной энтропии, ее связь с показателями качества сжатия и методы ее подсчета, выполнено сравнение операционных систем, проведен обзор загружаемых модулей ядра Linux, предоставляющих возможность сжатия оперативной памяти.

В разделе 2 изложены основные этапы оптимизации метода сжатия страниц памяти и метода подсчета информационной энтропии, спроектированы структура и тестирование программного обеспечения, реализующего оптимизацию метода.

В разделе 3 разработано и протестировано спроектированное программное обеспечение, описана его конфигурация.

В разделе 4 проведены исследование корреляции информационной энтропии и показателей качества сжатия, исследование соотношения времени сжатия и времени вычисления информационной энтропии и исследование зависимости показателей качества сжатия и времени сжатия от типов хранимых в памяти данных.

Ключевые слова: сжатие оперативной памяти, коэффициент сжатия, информационная энтропия, метод скользящего окна, биномиальный метод.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	8
1 Аналитический раздел	10
1.1 Управление памятью в операционной системе	10
1.1.1 Виртуальная память	10
1.1.2 Подкачка страниц	11
1.2 Сжатие данных	12
1.2.1 Сжатие памяти	13
1.2.2 Показатели качества сжатия	15
1.3 Информационная энтропия	16
1.3.1 Свойства информационной энтропии	17
1.3.2 Связь со сжатием данных	18
1.4 Методы подсчета информационной энтропии	18
1.4.1 Метод скользящего окна	18
1.4.2 Биномиальный метод	20
1.4.3 Сравнение существующих методов подсчета	22
1.5 Постановка задачи	23
1.6 Сравнение операционных систем	24
1.7 Сжатие памяти в ядре Linux	25
1.7.1 Модуль zram	25
1.7.2 Модуль zswap	27
2 Конструкторский раздел	29
2.1 Функциональная модель оптимизации метода сжатия страниц	29
2.2 Требования к разрабатываемому программному обеспечению	30
2.3 Структура разрабатываемого программного обеспечения	30
2.4 Выбор типов и структур данных	32
2.5 Описание оптимизации метода сжатия страниц	32
2.6 Проектирование тестирования программного обеспечения	34
3 Технологический раздел	38
3.1 Выбор средств программной реализации	38
3.2 Детали реализации	38

3.3	Конфигурация программного обеспечения	39
3.4	Тестирование разработанного программного обеспечения	41
4	Исследовательский раздел	45
4.1	Исследование корреляции энтропии и показателей качества сжа- тия	45
4.2	Исследование зависимости характеристик от типов хранимых в памяти данных	47
4.2.1	Исследование временной эффективности	49
	ЗАКЛЮЧЕНИЕ	51
	СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	55
	ПРИЛОЖЕНИЕ А	56

ВВЕДЕНИЕ

Основным ресурсом в современном обществе является информация [1]. Для информационного общества характерен процесс компьютеризации, при котором задачи различных сфер жизни решаются с использованием электронно-вычислительной техники. Управление ресурсами электронно-вычислительной машины является задачей операционной системы.

На производительность операционной системы влияет объем доступной оперативной памяти. Одним из способов увеличения оперативной памяти является сжатие данных [2]. В целях оптимизации необходимо оценивать показатели его качества, что может быть реализовано с помощью информационной энтропии.

Целью данной работы является разработка оптимизации метода сжатия страниц памяти с использованием подсчета информационной энтропии.

Для достижения поставленной цели необходимо выполнить следующие задачи:

- рассмотреть основные понятия управления памятью в операционной системе;
- провести анализ предметной области сжатия данных,
- изучить свойства информационной энтропии и ее связь со сжатием данных;
- описать известные методы подсчета информационной энтропии;
- сформулировать основные этапы оптимизации метода сжатия страниц памяти с использованием подсчета информационной энтропии;
- разработать программное обеспечение, реализующее данную оптимизацию метода;
- протестировать разработанное программное обеспечение;
- исследовать корреляцию информационной энтропии и показателей качества сжатия;

- установить зависимости времени сжатия и коэффициента сжатия от различных типов хранимых в памяти данных;
- исследовать временную эффективность разработанной оптимизации.

1 Аналитический раздел

1.1 Управление памятью в операционной системе

Для хранения данных запущенных в операционной системе процессов, находящихся в состоянии выполнения, ожидания или блокировки, используется оперативная память. Устройство, предназначенное для записи, считывания и хранения инструкций и данных процессов, называется оперативным запоминающим устройством (ОЗУ) [3]. В многозадачных системах память должна быть распределена для размещения нескольких процессов. Для управления памятью в системе используется абстракция адресного пространства. У каждого процесса имеется собственное адресное пространство — набор адресов, который может быть использован процессом для обращения к памяти [4]. Если объем оперативной памяти, необходимый для размещения данных всех процессов, превышает объем ОЗУ, то возникает перегрузка памяти. Одним из способов решения проблемы является виртуальная память.

1.1.1 Виртуальная память

Механизм виртуальной памяти предполагает разделение логической памяти и физической памяти. У каждого процесса имеется собственное виртуальное адресное пространство. При использовании страничной организации памяти, виртуальное адресное пространство делится на блоки заданного размера, называемые страницами. Физическая память делится на блоки заданного размера, называемые страничным блоком или кадром. Страницы виртуального адресного пространства и страничные блоки имеют одинаковый размер, который определяется аппаратным обеспечением. Значение размера представляет собой степень двойки и изменяется от четырех килобайт до одного гигабайта [5].

При использовании виртуальной памяти процессы оперируют виртуальными адресами. Виртуальный адрес — это адрес, присвоенный местоположению в виртуальной памяти, который позволяет обращаться к данному местоположению так, как если бы это была часть физической памяти [4]. Отображение виртуального адреса на физический адрес памяти выполняется диспетчером памяти с использованием таблиц страниц, как показано на рисунке 1.1.

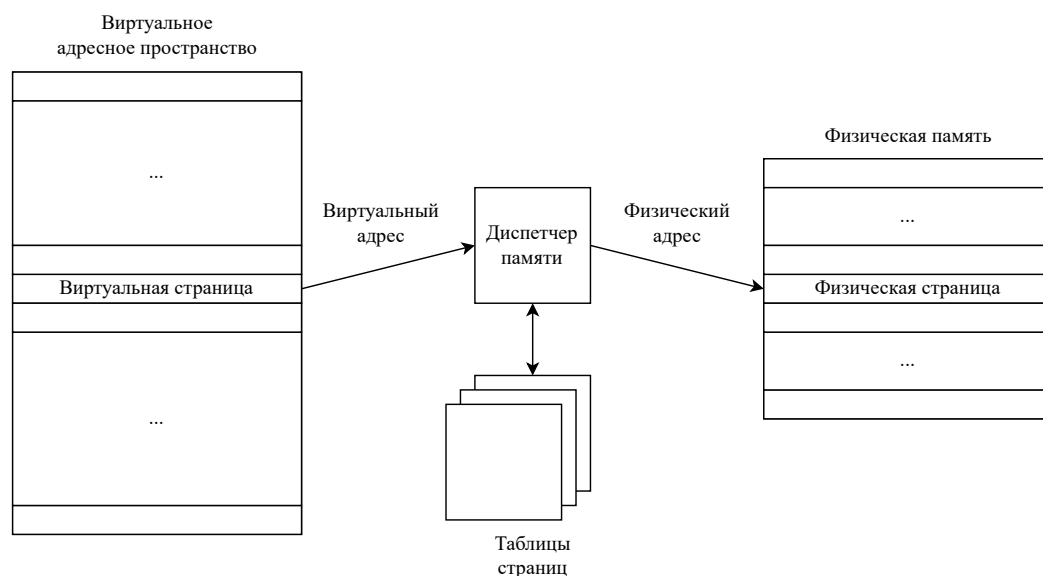


Рисунок 1.1 – Схема работы виртуальной памяти

Одним из основных преимуществ использования виртуальной памяти является то, что объем памяти, размещающий данные процессов не ограничен объемом доступной физической памяти. В системах с виртуальной памятью используется подкачка страниц по запросу.

1.1.2 Подкачка страниц

Данные процесса или их часть можно выгружать из оперативной памяти в резервное хранилище, а затем возвращать в память для продолжения выполнения. Резервное хранилище представляет собой вторичное хранилище, разделенное на блоки определенного размера, которые имеют тот же размер, что и страничные кадры. Таким образом, неиспользуемая страница памяти может быть перемещена в резервное хранилище, и загружена в оперативную память при необходимости. Этот принцип работы получил название подкачка страниц, при этом вторичное хранилище называется устройством подкачки, а его раздел, используемый для хранения данных — пространством подкачки [5].

Присутствие страницы в оперативной памяти отслеживается с помощью бита присутствия-отсутствия [4]. Если процесс попытается получить доступ к отсутствующей в памяти странице, возникает системное прерывание (page fault). Операционная система находит свободный страничный кадр или выбирает редко используемую страницу и сбрасывает ее содержимое во вторичное

хранилище. Затем она перемещает нужную физическую страницу в свободный страничный кадр. Схема работы подкачки страниц при возникновении page fault приведена на рисунке 1.2.

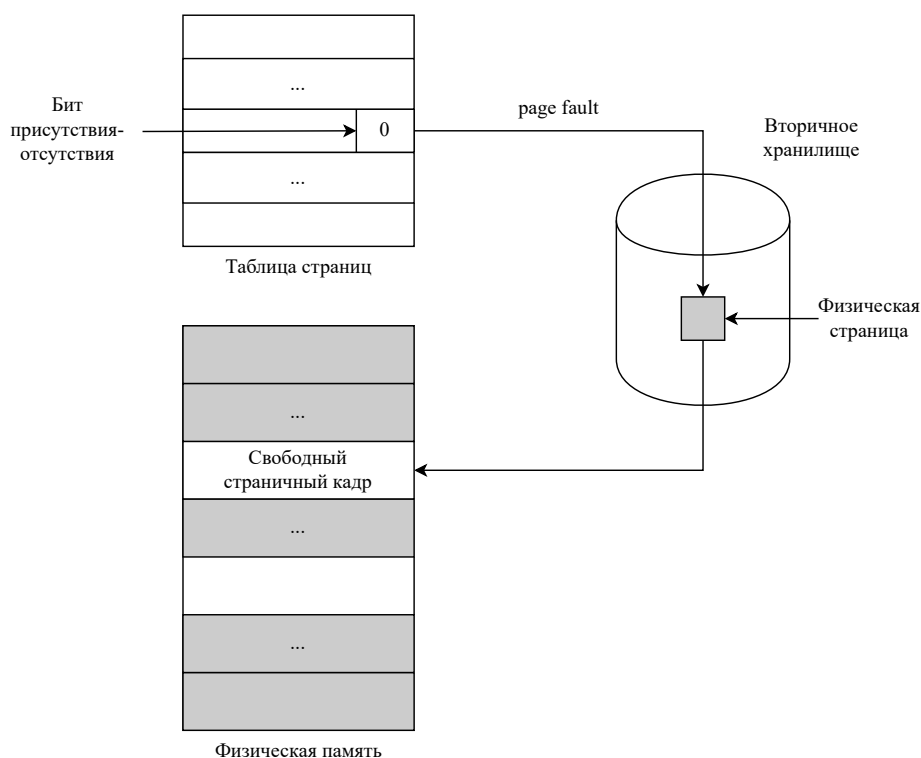


Рисунок 1.2 – Схема работы подкачки страниц при возникновении page fault

Подкачка страниц позволяет загружать только требуемые для выполнения данные, но для переноса страниц между вторичным хранилищем и оперативной памятью требуется время. Частое перемещение снижает производительность работы подсистемы памяти. Для решения этой проблемы используется сжатие памяти [2].

1.2 Сжатие данных

Неслучайные данные имеют некоторую структуру. Наличие у данных структуры, которую можно использовать для уменьшения их размера путем достижения такого представления данных, в котором никакая структура не выделяется, называют избыточностью. Сжатие данных — это процесс преобразования исходных данных в их компактную форму путем распознавания и использования избыточности данных [6]. Процесс сжатия состоит из двух этапов:

1. Этап моделирования, который включает в себя распознавание избыточности для построения модели. Модель представляет собой набор данных и правил, используемых для обработки входных символов.
2. Этап кодирования данных с использованием модели.

Описанные этапы сжатия данных представлены на рисунке 1.3.

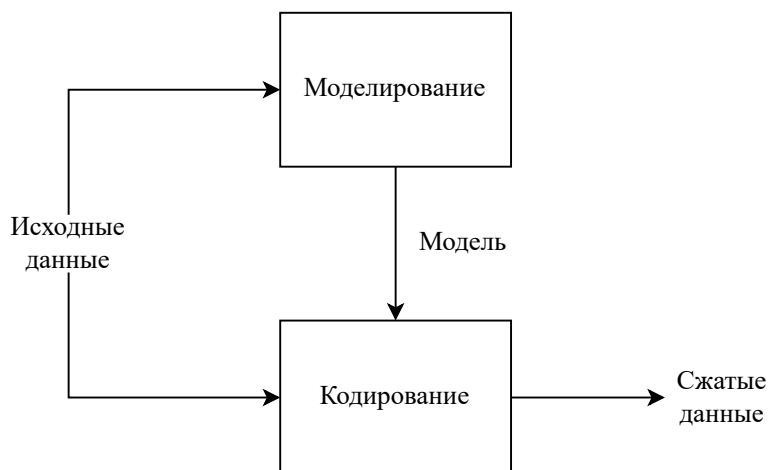


Рисунок 1.3 – Этапы сжатия данных

В результате сжатия исходных данных X получается их представление $X_{\text{сж}}$. При восстановлении сжатые данные $X_{\text{сж}}$ преобразуются в представление Y . На основании требований к восстановлению выделяют [7]:

- сжатие данных без потерь, при котором $Y = X$ (обратимое сжатие);
- сжатие данных с потерями, при котором $Y \neq X$ (необратимое сжатие).

Схемы обратимого и необратимого сжатия данных показаны на рисунке 1.4.

1.2.1 Сжатие памяти

Сжатие памяти заключается в том, что несколько страничных кадров сжимаются, и их сжатые версии сохраняются в меньшем числе страничных кадров. Это позволяет системе сократить использование памяти, не прибегая к подкачке страниц. Если процессу требуются сжатые страницы, они восстанавливаются [5].

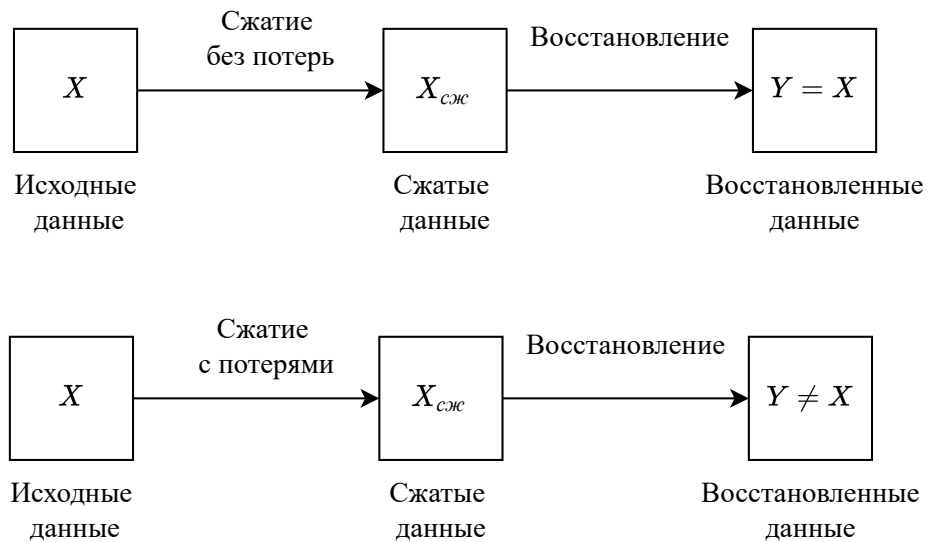


Рисунок 1.4 – Схемы обратимого и необратимого сжатия данных

На рисунке 1.5 приведена схема сжатия памяти: неиспользуемые страницы 2-5 хранятся в сжатом виде, в результате увеличивается объем свободной памяти.

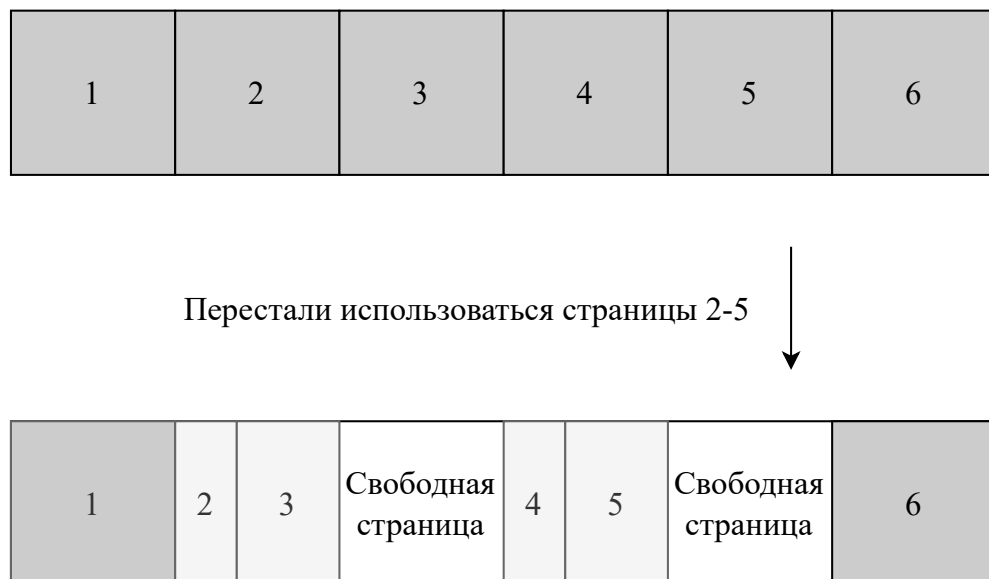


Рисунок 1.5 – Схема сжатия памяти

Выделяют следующие требования к реализации сжатия памяти [8]:

- при запросе на чтение данные должны восстанавливаться без потерь;
- операции сжатия и восстановления должны иметь низкую временную задержку и высокое качество сжатия.

1.2.2 Показатели качества сжатия

Для измерения качества сжатия используют следующие характеристики: коэффициент сжатия, сила сжатия (compression gain) и экономия пространства (space savings).

Коэффициент сжатия определяет средний объем памяти, необходимый для хранения сжатых данных, и вычисляется по формуле 1.1.

$$K_{\text{сж}} = \frac{L_{\text{исх}}}{L_{\text{сж}}}, \quad (1.1)$$

где $L_{\text{исх}}$ — объем исходных данных, $L_{\text{сж}}$ — объем сжатых данных. Высокий коэффициент сжатия говорит о высоком качестве сжатия. У сжатия низкого качества могут быть следующие причины [9]:

- исходные данные являются сжатыми;
- исходные данные являются зашифрованными;
- исходные данные не содержат повторяющихся шаблонов.

Сила сжатия определяется по формуле 1.2.

$$CG = 100 \cdot \log_e \frac{L_{\text{исх}}}{L_{\text{сж}}} = 100 \cdot \log_e K_{\text{сж}}. \quad (1.2)$$

Экономия пространства показывает уменьшение размера файла по сравнению с размером несжатого файла. Ее можно получить с помощью формулы 1.3.

$$SS = 1 - \frac{L_{\text{сж}}}{L_{\text{исх}}}. \quad (1.3)$$

Например, если размер исходных данных равен десяти мегабайтам, а размер сжатых данных равен трем мегабайтам, то экономия пространства указывает на то, что 70% пространства сохраняется благодаря сжатию.

Оценить показатели качества сжатия можно с помощью информационной энтропии.

1.3 Информационная энтропия

Под информацией понимают сведения, которые являются объектом хранения, передачи и обработки [10]. Формой представления информации является сообщение. Физическую величину, отображающую сообщение, называют сигналом. Передача информации осуществляется следующим образом [11]:

1. Источник информации создает случайное сообщение. В теории информации любой источник информации является стохастическим, его можно описать измеряемыми вероятностными категориям.
2. Сообщение поступает в систему передачи, в которой выполняется кодирование — преобразование сообщения с целью согласования источника информации с каналом связи для увеличения скорости передачи информации или обеспечения заданной помехоустойчивости [10]. Кодирование состоит из шифрования, сжатия и защиты от шума, в результате которых формируется сигнал.
3. Сигнал проходит через канал — среду передачи информации [11]. В канале могут возникать помехи, создаваемые источником шума.
4. Сигнал подается на вход системе приема, которая выполняет декодирование — восстановление исходного сообщения [10].
5. Исходное сообщение передается получателю.

Описанная схема передачи информации представлена на рисунке 1.6.

Сообщение содержит сведения о некоторой физической системе X , которая случайным образом может перейти в какое-либо состояние x_i из конечного множества состояний x_1, x_2, \dots, x_n с вероятностями p_1, p_2, \dots, p_n , где $n \in \mathbb{N}$, $p_i = P(X \sim x_i)$ и $\sum_{i=1}^n p_i = 1$. То есть, для такой системы существует степень неопределенности, которая описывается числом ее возможных состояний и их вероятностями. Сведения из принятого сообщения тем ценнее, чем больше была неопределенность системы до получения сообщения. Специальную характеристику, используемую в качестве меры неопределенности системы, называют информационной энтропией [12]. Информационная энтропия конечной вероятностной схемы определяется по формуле 1.4, которая называется формулой Шеннона.

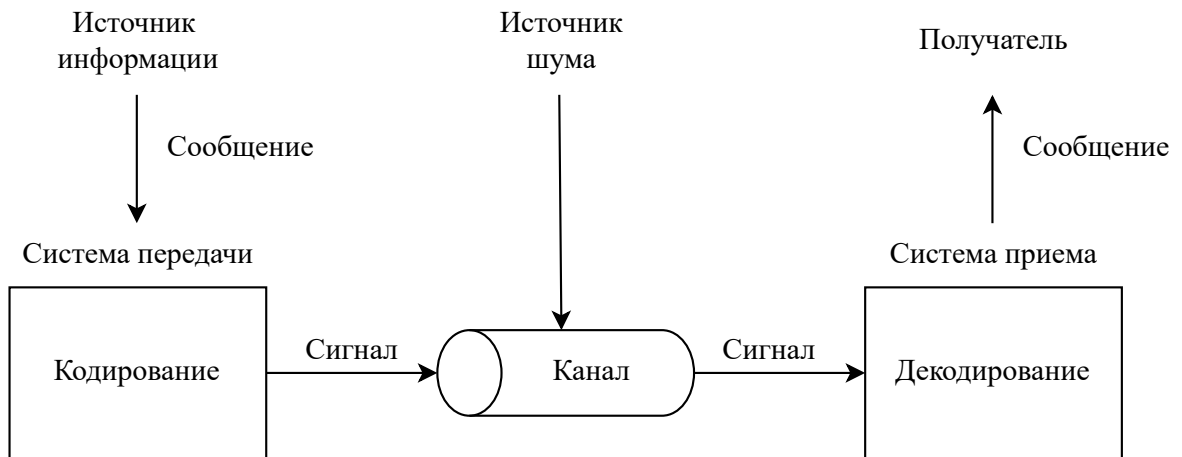


Рисунок 1.6 – Схема передачи информации

$$H(X) = - \sum_{i=1}^n (p_i \cdot \log_a p_i), \quad (1.4)$$

где $p_i \in [0, 1]$, $a > 1$.

Основание логарифма определяет единицы измерения информационной энтропии: при $a = 2$ энтропия измеряется в битах, при $a = 3$ — в тритах, при $a = e$ — в натах.

1.3.1 Свойства информационной энтропии

Информационная энтропия обладает следующими свойствами [13]:

1. Энтропия всегда неотрицательна. Значения $\log_a p_i$ в формуле (1.4) принимают неположительные значения, так как $p_i \in [0, 1]$. Поэтому верно неравенство $H(X) \geq 0$.
2. Энтропия равна нулю, если состояние системы в точности известно заранее. Если известно состояние x_k , в которое перейдет система X , то вероятность этого состояния p_k равна единице, вероятности других состояний равны нулю. Тогда выполняется равенство $p_k \cdot \log_a p_k = 0$. В связи с тем, что $\lim_{p \rightarrow 0} (p \cdot \log_a p) = 0$, другие слагаемые суммы в формуле (1.4) равны нулю. В этом случае верно равенство $H(X) = 0$.

3. Энтропия принимает наибольшее значение при условии, что все состояния равновероятны, то есть, $p_1 = p_2 = \dots = p_n = \frac{1}{n}$. Тогда выполняется равенство $H(X) = \log_a n$.

1.3.2 Связь со сжатием данных

Данные, обладающие предсказуемой структурой, сокращают неопределенность системы меньше, чем сведения, в которых никакая структура не выделяется. Так как информационная энтропия является мерой неопределенности системы, то данные с выделяемой структурой, имеют низкое значение энтропии. Сведения, в которых закономерности не определяются, имеют высокое значение энтропии [14]. Так, чем меньше избыточность данных, тем выше значение их энтропии. То есть, информационная энтропия сжатых данных выше, чем ее значение до сжатия.

Согласно теореме Шеннона о кодировании источника шифрования для источника сообщений с энтропией H невозможно сжать данные без потери информации так, что среднее число бит на символ будет меньше, чем H . Это означает, что невозможно сжать сообщение без потери данных до размера меньше, чем энтропия источника. Таким образом, на основании информационной энтропии исходных данных определяется теоретическая граница показателей качества сжатия [15].

В связи с применением операций сложения, умножения и логарифмирования при вычислении энтропии по формуле (1.4), число которых растет с увеличением объема данных, встает задача выбора метода подсчета. Использование метода влияет на скорость и время определения информационной энтропии.

1.4 Методы подсчета информационной энтропии

1.4.1 Метод скользящего окна

При решении задач, связанных со сжатием, данные X представляют собой массив байтов размером N [16]. Байт состоит из восьми битов, каждый из которых кодирует одно из значений множества $\{0, 1\}$. Поэтому один байт может принимать значения из интервала от 0 до 255 включительно в десятичной системе счисления.

В методе скользящего окна под окном понимают рассматриваемую на текущем этапе подпоследовательность данных размером n [17]. При подсчете энтропии данным методом необходима дополнительная память размером 2^n для хранения числа вхождений подпоследовательностей данных. Так как с увеличением размера окна, растет объем дополнительной памяти, в качестве окна выбирается минимально адресуемая единица памяти, которой является байт [18]. В связи с тем, что на каждом этапе окно смещается на следующие восемь битов, как показано на рисунке 1.7, оно называется скользящим.

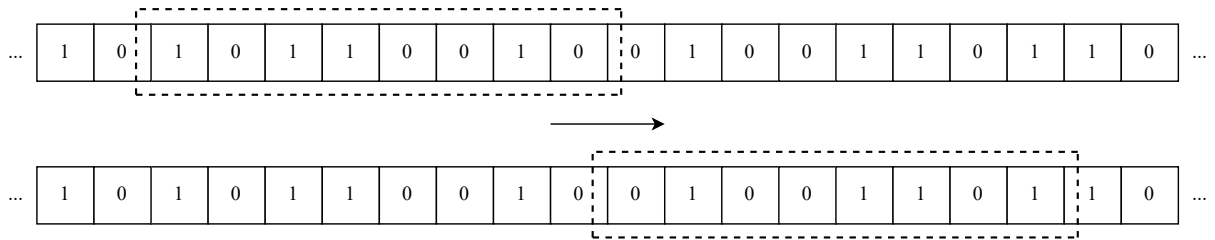


Рисунок 1.7 – Проход по массиву байтов методом скользящего окна

Вычисление информационной энтропии методом скользящего окна включает в себя два шага:

1. Для каждого возможного значения байта подсчитывается число его вхождений k_i в массив байтов, где $i = \overline{0, 255}$.
2. С учетом того, что вероятность появления байта в массиве $p_i = \frac{k_i}{N}$, информационная энтропия вычисляется по формуле 1.5.

$$H(X) = - \sum_{i=0}^{255} (p_i \cdot \log_2 p_i). \quad (1.5)$$

Так как первый шаг метода предполагает проход по массиву байтов размером N , а второй шаг — проход по массиву числа вхождений подпоследовательностей данных размером 2^n , временная сложность метода скользящего окна — $O(N + 2^n)$.

Информационная энтропия, подсчитанная методом скользящего окна, принимает значения из интервала $[0, 8]$ битов. При этом согласно свойству 2 из подраздела 1.3.1 информационная энтропия принимает нулевое значение в случае, когда массив данных состоит из одинаковых байтов, и в соответствии

со свойством 3 из подраздела 1.3.1 максимальное значение, равное восьми, если все байты в массиве различны.

1.4.2 Биномиальный метод

В биномиальном методе вычисления информационной энтропии [19] данные X рассматриваются как последовательность сообщений, генерируемых бернуллиевским источником — источником информации, порождающим символы из алфавита $\{0; 1\}$ с вероятностями $1 - p$ и p соответственно, причем $p \in (0, 1)$ и может быть неизвестно [20]. То есть, сообщение представляет собой последовательность битов длины n .

Вероятность того, что сообщение содержит k единиц, где $k = \overline{0, n}$, вычисляется по формуле 1.6.

$$P_k = p^k \cdot (1 - p)^{(n-k)}. \quad (1.6)$$

Количество возможных сообщений, содержащих k единиц, определяется как биномиальный коэффициент с использованием формулы 1.7.

$$C_n^k = \frac{n!}{k! \cdot (n - k)!}. \quad (1.7)$$

В связи с тем, что k может принимать значения из интервала $[0, n]$, число биномиальных коэффициентов, подсчитанных по формуле (1.7), равно $n + 1$. Это означает, что сообщения можно разбить на $n + 1$ классов эквивалентности. Тогда информационная энтропия вычисляется по формуле 1.8.

$$H(X) = - \sum_{k=0}^n (C_n^k \cdot P_k \cdot \log_2 P_k). \quad (1.8)$$

В соответствии с формулой (1.8) биномиальный метод подсчета информационной энтропии состоит из следующих этапов:

1. Определяется вероятность p появления единицы в сообщениях.
2. Исходные сообщения разбиваются на $n + 1$ классов эквивалентности, сообщения которых содержат $k = \overline{0, n}$ единиц.
3. Для каждого класса эквивалентности рассчитывается биномиальный коэффициент C_n^k .

4. Вычисляются вероятности P_k по формуле (1.6).
5. Суммируются произведения биномиальных коэффициентов C_n^k , вероятностей P_k и логарифмов вероятностей P_k для всех $n + 1$ значений k .

Схема определения биномиальных коэффициентов C_n^k и вероятностей P_k приведена на рисунке 1.8.

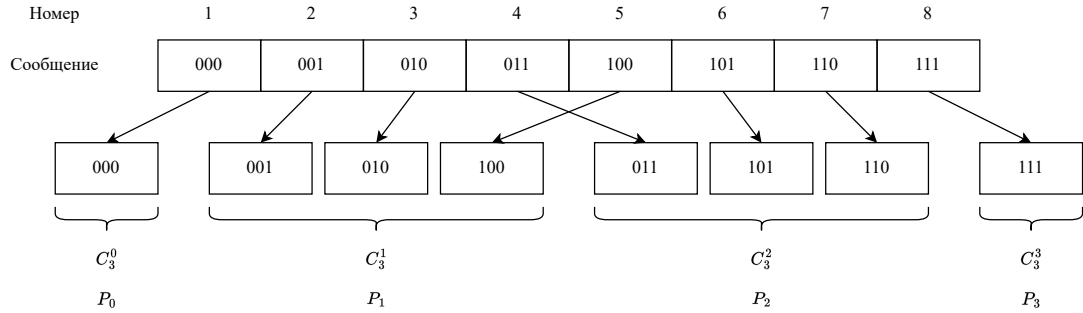


Рисунок 1.8 – Определение биномиальных коэффициентов C_n^k и вероятностей P_k

Этап определения вероятности p появления единицы в сообщениях подразумевает проход по массиву байтов размером N , в котором с единицей сравнивается каждый бит одного байта. Последний этап состоит из $n + 1$ цикла вычисления. При определении вероятности того, что сообщение содержит k единиц, требуются возведение в степень k вероятности появления единицы p и возведение в степень $n - k$ вероятности появления нуля $1 - p$. Тогда временная сложность биномиального метода подсчета информационной энтропии — $O(N + n^2)$.

В биномиальном методе определяются вероятности появления не подпоследовательности, а единицы в подпоследовательностях. За счет разделения исходной последовательности на классы эквивалентности сокращается число операций сложения.

При рассмотрении бернуллиевского источника информации предполагается, что вероятность появления единицы в подпоследовательности не зависит от вероятностей появления нуля или единицы в битах предыдущей подпоследовательности. При наличии такой зависимости вычисленное значение энтропии будет завышено. Так как рассматриваемое представление данных —

последовательность битов, то вероятности появления каждого значения бита независимы.

Недостатком данного метода является необходимость вычисления факториала при определении биномиальных коэффициентов и возведения в степень при определении вероятностей появления сообщений. Для снижения времени подсчета биномиальных коэффициентов можно хранить их значения в дополнительном массиве. Так как для сообщения длины n количество требуемых для определения энтропии биномиальных коэффициентов равно $n + 1$, то для их хранения потребуется дополнительная память размером $n + 1$.

1.4.3 Сравнение существующих методов подсчета

Для сравнения методов подсчета информационной энтропии были выделены следующие критерии оценки:

- К1 — временная сложность;
- К2 — использование базовой вычислительной арифметики;
- К3 — возможность распараллеливания вычислений;
- К4 — сложность по памяти.

Результаты сравнения представлены в таблице 3.1.

Таблица 1.1 – Сравнение методов подсчета информационной энтропии

Метод	К1	К2	К3	К4
Скользящего окна	$O(N + 2^n)$	+	+	2^n
Биномиальный	$O(N + n^2)$	–	+	n

Таким образом, биномиальный метод подсчета информационной энтропии требует меньших вычислительных затрат по времени и по памяти, чем метод скользящего окна, но использует трудоемкие вычислительные операции. При решении задачи оценки качества сжатия помимо времени подсчета информационной энтропии важна корреляция полученного значения и показателей качества сжатия. При этом не так важна точность определяемого значения.

1.5 Постановка задачи

Из проведенного в подразделах 1.2-1.3 анализа сжатия и информационной энтропии можно сделать следующие выводы:

- сжатие и восстановление страниц оперативной памяти должно иметь низкую временную задержку и высокое качество сжатия;
- сжатие может обладать низким качеством, если исходные данные являются сжатыми, зашифрованными или не содержат повторяющихся шаблонов;
- значение информационной энтропии исходных данных позволяет определить теоретическую границу показателей качества сжатия.

Так, сжатие может иметь низкое качество, при этом на него будет тратиться время. Можно сократить временные затраты на сжатие, если определять теоретическую границу качества сжатия до его выполнения, и на основании оценки качества принимать решение, сжимать страницу или нет. Качество сжатия по исходным данным можно оценить с помощью информационной энтропии. Таким образом, необходимо провести оптимизацию сжатия страниц памяти, принимая решение о его выполнении на основании вычисленного значения информационной энтропии исходных данных.

Входными данными поставленной задачи является страница памяти, которая представляет собой вектор $a = (a_1 \ a_2 \ \dots \ a_N)$ размером N , равным размеру страницы памяти. При этом, $0 \leq a_i \leq 255$.

На принятие решения о выполнении сжатия влияет сравнение подсчитанного значения информационной энтропии с пороговым значением. Сжатие страниц памяти выполняется ядром операционной системы.

В зависимости от принятого решения выходными данными могут быть:

- страница памяти со сжатыми исходными данными, которая представляет собой вектор $b = (b_1 \ b_2 \ \dots \ b_N)$ размером N , $0 \leq b_i \leq 255$;
- страница памяти с исходными данными, которая представляет собой вектор a .

На рисунке 1.9 показана IDEF0-диаграмма нулевого уровня, формализующая поставленную задачу.

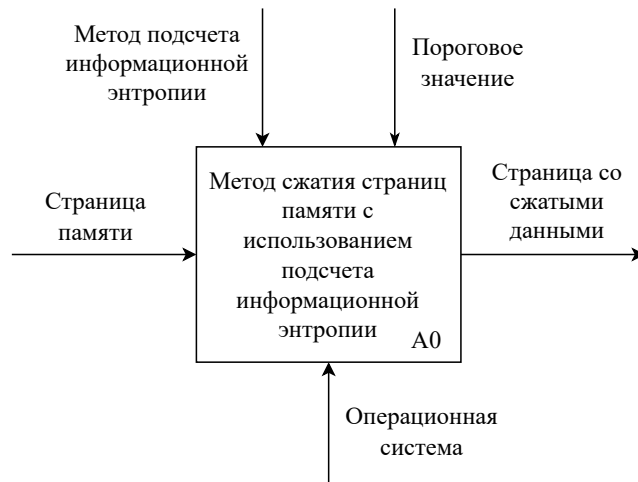


Рисунок 1.9 – IDEF0-диаграмма нулевого уровня

1.6 Сравнение операционных систем

Выполнение оптимизации метода сжатия страниц памяти должно проводиться на операционной системе, удовлетворяющей следующим критериям:

- поддержка сжатия страниц оперативной памяти [5];
- открытый исходный код [21].

Для сравнения, приведенного в таблице 1.2, были выбраны системы, занимающие наибольшие доли рынка операционных систем [22].

Таблица 1.2 – Сравнение операционных систем

Операционная система	Поддержка сжатия	Открытый исходный код	Доля рынка, %
Windows	+	-	63.13
macOS	+	-	17.78
Linux	+	+	2.83
FreeBSD	+	+	0.01

В дальнейшем будет рассматриваться ядро операционной системы Linux, так как она удовлетворяет выделенным критериям и занимает долю рынка большую, чем операционная система FreeBSD.

1.7 Сжатие памяти в ядре Linux

Ядро Linux является модульным. Модуль — это часть кода, которую можно загрузить и выгрузить в ядро по необходимости [23]. Модуль может быть встроенным или загружаемым.

Код встроенного модуля встраивается в образ ядра во время компиляции. Если ядру потребуется функциональность этого модуля, оно инициализирует и загрузит модуль. Для того, чтобы изменения такого модуля вступили в силу, необходимо перекомпилировать ядро.

Загружаемый модуль может быть загружен пользователем в ядро во время работы системы. Такие модули используют для добавления требуемой функциональности ядра без перезагрузки системы. Это позволяет избежать пересборки и перезагрузки ядра при каждом добавлении новых функций.

Возможность сжатия страниц оперативной памяти в ядре Linux предоставляют загружаемые модули `zram` и `zswap`.

1.7.1 Модуль `zram`

При загрузке модуля `zram` в оперативной памяти создается блочный специальный файл. Специальные файлы в операционной системе Linux служат для реализации работы с устройствами ввода-вывода как с файлами и находятся в каталоге `/dev`. Это позволяет драйверам устройств проводить с устройствами ввода-вывода операции чтения и записи, используя те же системные вызовы, которые применяются для чтения и записи файлов. Блочные специальные файлы используются для моделирования устройств, содержащих набор блоков с произвольной адресацией. Открыв такой файл, программа может получить доступ к любому блоку устройства независимо от структуры имеющейся файловой системы [24]. При этом данные могут быть прочитаны или записаны только кратными блоками.

Созданный модулем `zram` блочный специальный файл моделирует диск. После загрузки модуля страницы, попадаемые на вход моделируемому диску, сжимаются и хранятся сжатыми. Таким образом, страницы находятся в

оперативной памяти в сжатом виде.

На рисунке 1.10 представлена схема работы модуля zram.

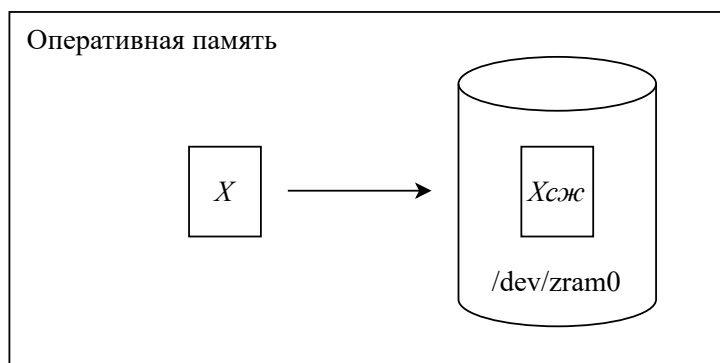


Рисунок 1.10 – Схема работы модуля zram

Модуль zram дает следующие преимущества [25]:

- сокращение использования памяти за счет сжатия страниц;
- высокая скорость операций чтения и записи, так как блочный специальный файл создается в оперативной памяти;
- возможность выбора алгоритма сжатия, предоставляемого Crypto API [26];
- возможность реализации хранения временных файлов в сжатом виде;
- возможность реализации кэширования страниц памяти;
- возможность использования в качестве устройства подкачки;
- возможность записи несжимаемых страниц в резервное хранилище;
- возможность повторного сжатия с другим алгоритмом сжатия;
- многопоточное сжатие.

1.7.2 Модуль zswap

Модуль zswap работает следующим образом:

- перехватывает страницу, находящуюся в процессе выгрузки на устройство подкачки;
- осуществляет попытку записать данную страницу в сжатом виде в динамически создаваемый пул оперативной памяти;
- если размер пула сжатых страниц достигает предела, страницы вытесняются на устройство подкачки с использованием алгоритма LRU.

На рисунке 1.11 показана схема работы модуля zswap: при попытке записи страницы Z в заполненный пул длительно не используемая страница Y перемещается на устройство подкачки.

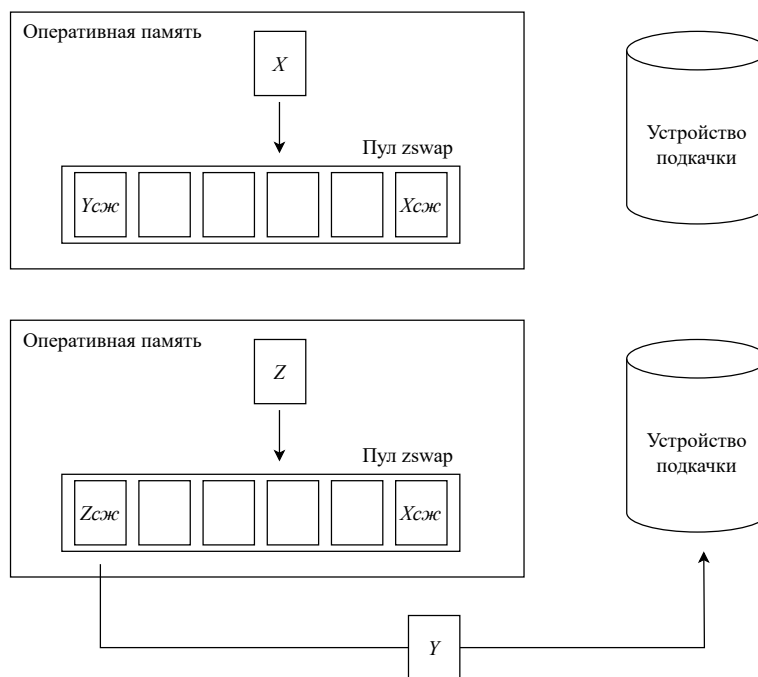


Рисунок 1.11 – Схема работы модуля zswap

Максимальный размер сжатого пула можно установить, изменяя значение параметра `max_pool_percent`.

Модуль zswap дает следующие преимущества [27]:

- сокращение использования памяти за счет сжатия страниц;

- высокая скорость операций чтения и записи, так как сжатый пул находится в оперативной памяти;
- возможность выбора алгоритма сжатия, предоставляемого Crypto API [26].

Вывод

В данном разделе были рассмотрены такие механизмы подсистемы памяти операционной системы, как виртуальная память и подкачка страниц. В связи с недостатками перемещения данных между вторичным хранилищем и оперативной памятью было изучено сжатие памяти, которое является альтернативой подкачки страниц. Были описаны методы подсчета информационной энтропии, с помощью которой можно оценить качество сжатия. В результате проведенного анализа была поставлена и формализована задача оптимизации метода сжатия страниц оперативной памяти с использованием подсчета информационной энтропии.

На основании выполненного сравнения операционных систем для решения поставленной задачи была выбрана операционная система Linux, поддерживающая сжатие страниц оперативной памяти и имеющая открытый исходный код. Был произведен обзор модулей сжатия в ядре Linux. В данной работе будет выполняться оптимизация сжатия страниц памяти загружаемым модулем `zgam` в связи с его преимуществами и тем, что для работы модуля `zswap` нужно устройство подкачки, на которое при полном заполнении пула вытесняются страницы памяти.

2 Конструкторский раздел

2.1 Функциональная модель оптимизации метода сжатия страниц

Разрабатываемая оптимизация метода сжатия страниц памяти с использованием подсчета информационной энтропии состоит из следующих этапов:

1. Вычисление значения информационной энтропии с помощью метода подсчета.
2. Сравнение вычисленного значения информационной энтропии с пороговым значением.
3. Сжатие данных страницы в случае, если полученное значение информационной энтропии меньше порогового значения.

IDEF0-диаграмма первого уровня, формализующая основные этапы оптимизации сжатия страниц оперативной памяти, приведена на рисунке 2.1.

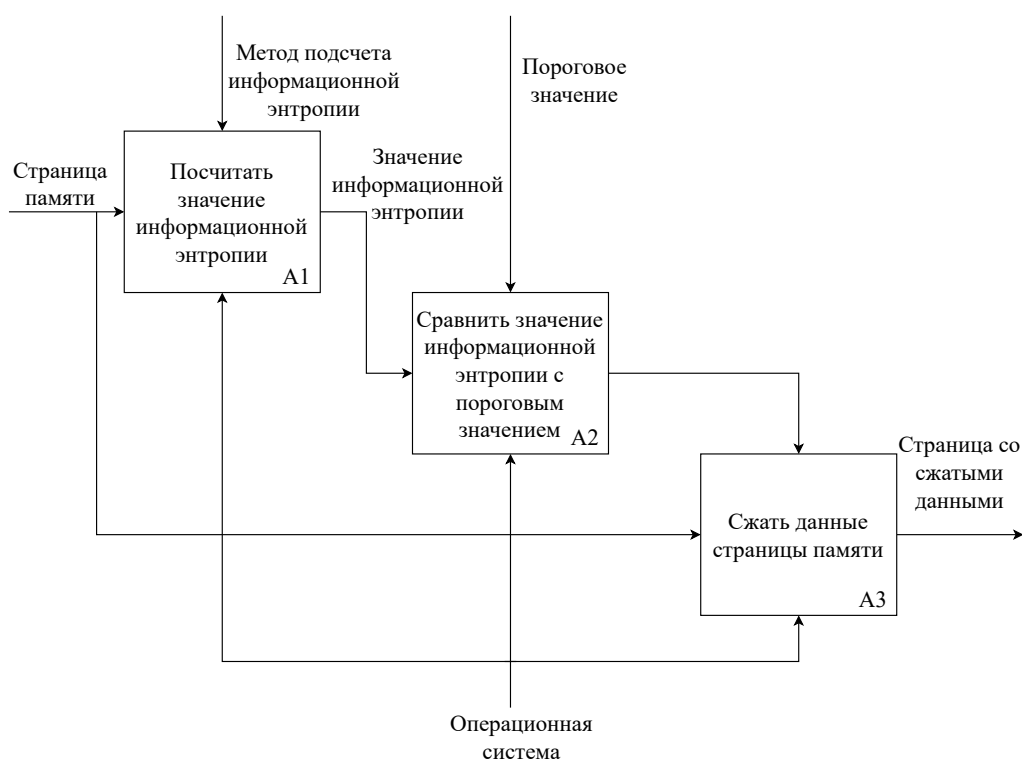


Рисунок 2.1 – IDEF0-диаграмма первого уровня

2.2 Требования к разрабатываемому программному обеспечению

Согласно описанию этапов решения поставленной задачи разрабатываемое программное обеспечение должно:

- вычислять информационную энтропию страницы оперативной памяти, которая является вектором $a = (a_1 \ a_2 \ \dots \ a_N)$ размером N , равным размеру страницы памяти, $0 \leq a_i \leq 255$;
- если вычисленное значение меньше порогового, сжимать данные входной страницы памяти, то есть, получать вектор $b = (b_1 \ b_2 \ \dots \ b_N)$ размером N , $0 \leq b_i \leq 255$;
- если вычисленное значение больше порогового или равно ему, данные входной страницы памяти не должны изменяться.

При этом пороговое значение должно устанавливаться на этапе сборки программного обеспечения.

2.3 Структура разрабатываемого программного обеспечения

Структура загружаемого модуля ядра `zgam` включает в себя следующие части:

- модуль блочного устройства, который выполняет функции создания, настройки и удаления дисков, обработки операций записи и чтения страниц и получения статистики;
- модуль сжатия, который вызывает функции сжатия и восстановления данных, а также настройку этих операций.

Функция сжатия данных, предоставляемая модулем сжатия, вызывается в модуле блочного устройства во время обработки записи страницы на диск, как представлено на рисунке 2.2.

Одним из выделенных к разрабатываемому программному обеспечению требованием является то, что сжатие страницы должно проводиться только в

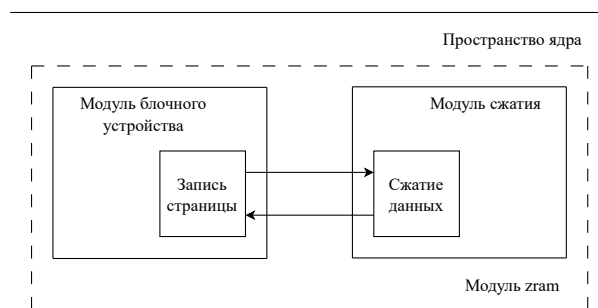


Рисунок 2.2 – Структура модуля zram

случае, если полученное значение информационной энтропии меньше порогового. Поэтому для выполнения поставленной задачи необходимо изменить функцию записи страницы на диск модуля блочного устройства. Построенная в соответствии с этим структура разрабатываемого программного обеспечения показана на рисунке 2.3. Пунктирными стрелками на схеме обозначены возможные переходы.

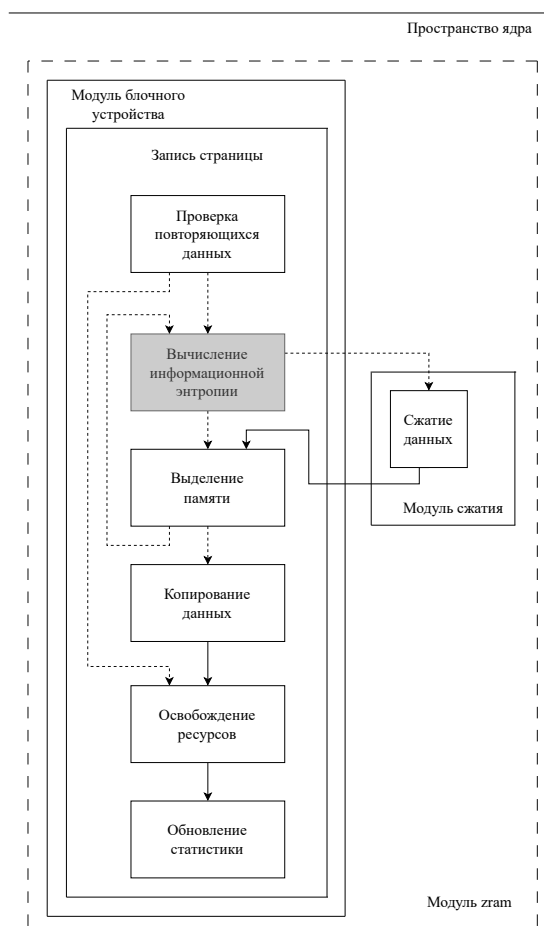


Рисунок 2.3 – Структура разрабатываемого программного обеспечения

2.4 Выбор типов и структур данных

Входными и выходными данными является страница памяти, которая задается вектором размером, меньшим или равным размеру страницы в байтах, состоящим из значений от нуля до 255. Поэтому для представления страницы в методе подсчета информационной энтропии должен использоваться массив типа `unsigned char` размером, равным размеру страницы в байтах.

При выполнении операций с числами с плавающей точкой в режиме пользователя ядро перехватывает системное прерывание и переходит из режима вычислений с целыми числами в режим с плавающей точкой. Если использовать режим с плавающей точкой в пространстве ядра, необходимо сохранять и восстанавливать состояние регистров с плавающей точкой математического сопроцессора. Вычисления с числами с плавающей точкой в режиме ядра выполнять не рекомендуется [28]. Поэтому для представления числовых величин должен использоваться целочисленный тип данных.

2.5 Описание оптимизации метода сжатия страниц

На разрабатываемый метод подсчета информационной энтропии накладываются следующие ограничения:

- вычисляемое значение может представлять неточное значение информационной энтропии, но должно обладать корреляцией с показателями качества сжатия;
- все числовые операнды в методе должны быть целыми числами.

В связи с тем, что при вычислении информационной энтропии биномиальным методом используются трудоемкие операции, разрабатываемый метод является модификацией метода скользящего окна. В методе скользящего окна информационная энтропия вычисляется по формуле 2.1.

$$H(X) = - \sum_{i=0}^{255} (p_i \cdot \log_2 p_i), \quad (2.1)$$

где k_i — число вхождений значения байта i в массив байтов, представляющий входную страницу, $p_i = \frac{k_i}{N}$ — вероятность появления байта в массиве

байтов, $i = \overline{0, 255}$, N — размер страницы в байтах. Тогда равенство 2.1 можно преобразовать в равенство 2.2.

$$H(X) = - \sum_{i=0}^{255} \left(\frac{k_i}{N} \cdot \log_2 \frac{k_i}{N} \right). \quad (2.2)$$

В соответствии со свойством логарифма частного верно равенство 2.3.

$$H(X) = - \sum_{i=0}^{255} \left(\frac{k_i}{N} \cdot (\log_2 k_i - \log_2 N) \right). \quad (2.3)$$

Равенство 2.3 эквивалентно равенству 2.4.

$$H(X) = \sum_{i=0}^{255} \left(\frac{k_i}{N} \cdot (\log_2 N - \log_2 k_i) \right). \quad (2.4)$$

Так как все числовые операнды являются целыми числами, при определении p_i выполняется целочисленное деление и вычисляемое значение округляется. Для сохранения большей точности результатом разрабатываемого метода будет являться значение информационной энтропии, умноженное на размер страницы. Для его определения используется формула 2.5.

$$H'(X) = \sum_{i=0}^{255} (k_i \cdot (\log_2 N - \log_2 k_i)), \quad (2.5)$$

Для сокращения времени вычисления логарифма числа вхождений значения байта i в массив байтов по основанию два в разрабатываемом методе используется дополнительный массив размером $\log_2 N$, для которого верно следующее:

- индекс элемента j является целым значением логарифма числа вхождений значения байта i , $j = \overline{\log_2 1, \log_2 N}$;
- значение элемента с индексом j является наибольшим целым числом x таким, что $\log_2 x = j$.

Для определения $\log_2 k_i$ необходимо взять индекс первого элемента описанного массива, который больше или равен k_i .

Схема алгоритма разрабатываемого метода подсчета информационной энтропии приведена на рисунке 2.4.

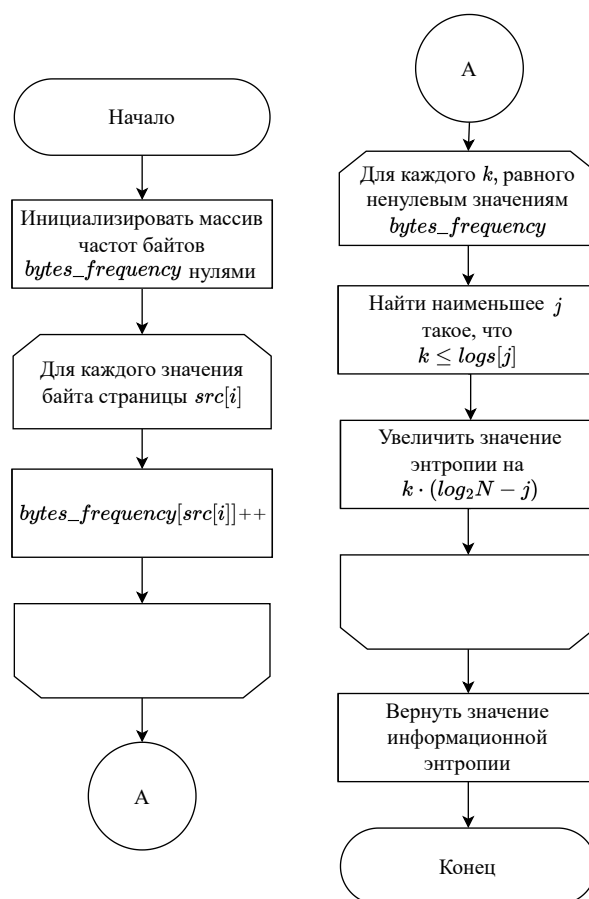


Рисунок 2.4 – Схема алгоритма подсчета информационной энтропии

В результате анализа исходного кода загружаемого модуля `zgam` [29] и описания этапов разрабатываемой оптимизации метода сжатия страниц памяти была построена схема алгоритма, представленная на рисунке 2.5.

2.6 Проектирование тестирования программного обеспечения

Разрабатываемое программное обеспечение выполняется при записи данных на диск `zgam`. Поэтому для тестирования программного решения необходимо моделировать процесс загрузки страниц памяти на диск. Проверка и отладка разрабатываемого программного обеспечения должны проводиться с помощью ручного тестирования.

Для проведения тестирования выделены следующие классы эквивалентности:

1. Несжимаемые страницы памяти;

2. Сжимаемые страницы памяти.

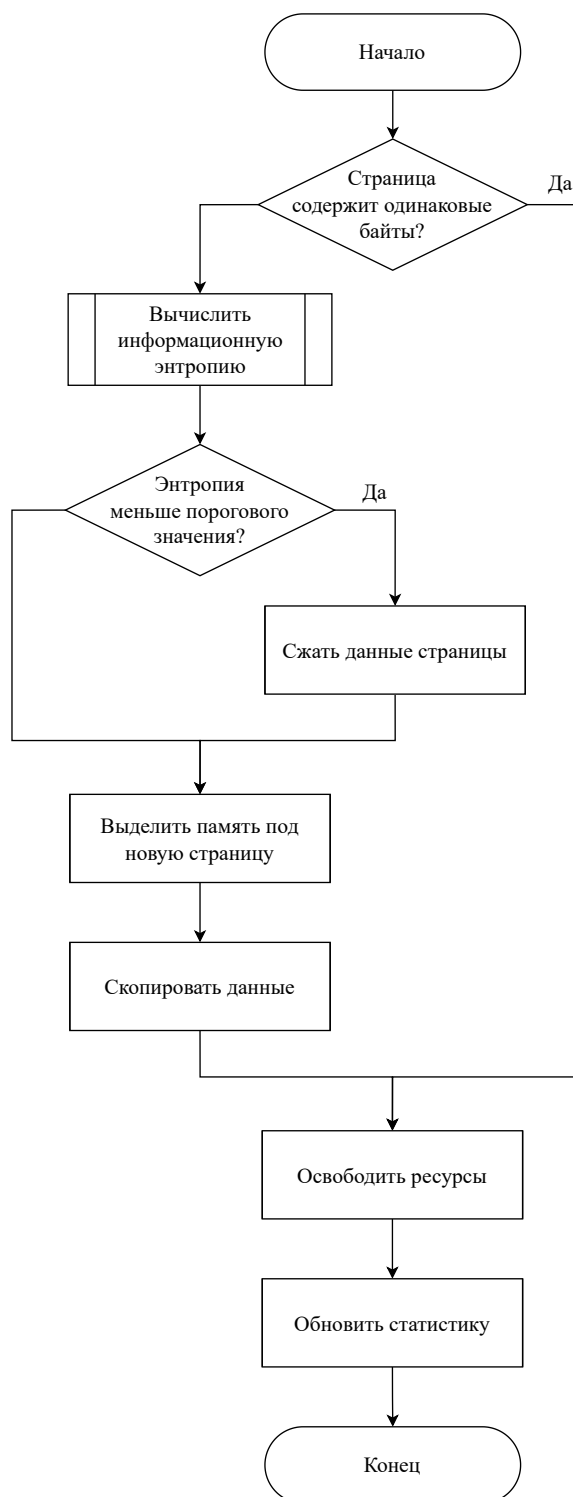


Рисунок 2.5 – Схема алгоритма записи страницы на диск

Несжимаемыми страницами памяти являются страницы с высоким значением информационной энтропии. В подразделе 1.3.2 было доказано, что

энтропия сжатых данных выше энтропии исходных данных. Поэтому тестовые наборы для первого класса эквивалентности должны включать сжатые данные. Примерами таких данных являются файлы форматов .jpg, .pdf, .mp3 и .zip [30].

Сжимаемыми страницами памяти являются страницы данных с выделяемой структурой, как было сказано в подразделе 1.3.2. Примерами таких данных являются файлы форматов .txt, .doc, .c, .h и исполняемые файлы [31], из которых могут состоять тестовые наборы для второго класса эквивалентности.

В результате каждого теста должна быть получена статистика сжатия тестового набора, включающая в себя следующую информацию:

- объем исходных данных в байтах;
- объем данных после сжатия в байтах;
- число несжимаемых страниц;
- число страниц с одинаковыми байтами.

Система для выполнения тестирования включает следующие компоненты:

- персональный компьютер, который является хостом;
- виртуальная машина, на которой загружен модуль zgam без разрабатываемой оптимизации;
- виртуальная машина, на которой загружен модуль zgam с разрабатываемой оптимизацией.

Тестирование разрабатываемого программного обеспечения состоит из следующих этапов:

1. Загрузить тестовый набор с хоста на диск zgam виртуальной машины с работающим без разрабатываемой оптимизации модулем zgam.
2. Получить статистику выполнения предыдущего шага.

3. Загрузить тестовый набор с хоста на диск zram виртуальной машины с работающим с разрабатываемой оптимизацией модулем zram.
4. Получить статистику выполнения предыдущего шага.
5. Сравнить полученные статистики.

Схема тестирующей системы представлена на рисунке 2.6.

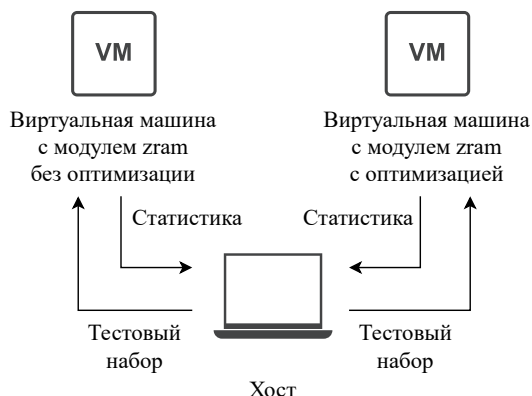


Рисунок 2.6 – Схема системы для тестирования

Вывод

В данном разделе были разработаны основные этапы оптимизации метода сжатия страниц памяти с использованием подсчета информационной энтропии. Их описание было приведено в виде псевдокода и схем алгоритмов. Были сформулированы требования к разрабатываемому программному решению. Взаимодействие компонентов системы было представлено в виде структуры программного обеспечения. Был обоснован выбор целочисленного типа данных и массива в качестве структуры, представляющей страницу памяти. Были выделены классы эквивалентности, описаны тестовые наборы для ручного тестирования, получаемая в результате статистика и тестирующая система.

3 Технологический раздел

3.1 Выбор средств программной реализации

Для выполнения оптимизации сжатия страниц памяти была выбрана операционная система Linux в соответствии со сравнением систем, проведенном в разделе 1.6. Согласно анализу модулей сжатия, представленному в разделе 1.7, был выбран загружаемый модуль zram. В качестве языка программирования был выбран язык C [32], так как большая часть исходного кода ядра операционной системы Linux и всех ее модулей написана на данном языке программирования.

Для сборки программного обеспечения выбрана утилита GNU make [33], так как с помощью данной утилиты осуществляется сборка загружаемых модулей ядра.

3.2 Детали реализации

В листинге 3.1 показана часть реализации записи страницы на диск в модуле блочного устройства. ENTROPY_LIMIT — константа, определяющая пороговое значение и устанавливаемая перед сборкой программного обеспечения.

Листинг 3.1 — Часть реализации записи страницы на диск

```
1  static int __zram_bvec_write(struct zram *zram,
2      struct bio_vec *bvec, u32 index, struct bio *bio)
3  {
4      ...
5      unsigned int comp_len = 0;
6      void *src, *dst, *mem;
7      struct zcomp_strm *zstrm;
8      struct page *page = bvec->bv_page;
9      ...
10     compress_again:
11         zstrm = zcomp_stream_get(zram->comp);
12         src = kmap_atomic(page);
13         if (get_entropy((const u8 *)src) < ENTROPY_LIMIT)
14             ret = zcomp_compress(zstrm, src, &comp_len);
15         else
16             comp_len = PAGE_SIZE;
17         ...
18     }
```

В листинге 3.2 приведена реализация разработанного метода подсчета информационной энтропии. Размер страницы в байтах в ядре Linux определяется константой `PAGE_SIZE` [24].

Листинг 3.2 — Реализация метода подсчета информационной энтропии

```
1  #define PAGE_SIZE_POW 12
2  #define BYTES_NUM 256
3
4  static inline s32 get_entropy(const u8 *src)
5  {
6      u16 logs[PAGE_SIZE_POW + 1] = {1, 3, 7, 15, 31, 63,
7          127, 255, 511, 1023, 2047, 4095, 4096};
8      u16 bytes_frequency[BYTES_NUM] = { 0 };
9      u32 i, j;
10
11     for (i = 0; i < PAGE_SIZE; ++i) {
12         bytes_frequency[src[i]]++;
13     }
14
15     s32 entropy = 0;
16
17     for (i = 0; i < BYTES_NUM; ++i) {
18         s32 k = bytes_frequency[i];
19
20         if (k > 0) {
21             j = 0;
22
23             while (k > logs[j])
24                 j++;
25
26             entropy += k * (PAGE_SIZE_POW - logs[j]);
27         }
28     }
29
30     return entropy;
31 }
```

3.3 Конфигурация программного обеспечения

Для сборки и запуска программного обеспечения необходимо выполнить следующие действия:

1. получить исполняемый файл программного обеспечения;
2. загрузить полученный модуль.

При загрузке модуля автоматически создается диск zram с идентификатором 0. Для корректной работы программного обеспечения нужно настроить созданный диск:

1. установить алгоритм сжатия из доступных в системе;
2. установить размер диска, для указания единиц измерения которого используются следующие постфиксы: К — размер в килобайтах, М — размер в мегабайт, G — размер в гигабайтах, размер в байтах устанавливается при отсутствии постфикса.

При получении статистики во время тестирования анализируются файл статистики памяти диска mm_stat.

Для остановки работы программного обеспечения необходимо выгрузить модуль. При этом автоматически удаляются все созданные диски zram.

Для выполнения при сборке, запуске, тестировании и остановке программного обеспечения и настройке диска описанных действий был написан make-файл, код которого показан в листинге 3.3. В таблице представлено описание доступных команд.

Листинг 3.3 — Конфигурационный файл

```
1  KERNEL_VERSION = $(shell uname -r)
2  MODULE_DIR = $(shell pwd)
3
4  OBJS = zcomp.o zram_drv.o
5  TARGET = zram
6
7  obj-m := $(TARGET).o
8  $(TARGET)-objs := $(OBJS)
9
10 all:
11     $(MAKE) -C /lib/modules/$(KERNEL_VERSION)/build M=$(MODULE_DIR) modules
12
13 $(TARGET).o: $(OBJS)
14     $(LD) -r -o $@ $(OBJS)
15
16 clean:
17     @rm -f *.o *.cmd *.flags *.mod.c *.order
18     @rm -f *.*.cmd *~ *.*~
19     @rm -fR .tmp*
20     @rm -rf .tmp_versions
21
```

Продолжение листинга 3.3

```
22  distclean:
23      @rm -f *.ko *.symvers *.mod
24
25  load:
26      sudo insmod $(TARGET).ko
27
28  info:
29      sudo lsmod | grep $(TARGET)
30
31  log:
32      sudo dmesg | grep $(TARGET)
33
34  unload:
35      sudo rmmod $(TARGET).ko
36
37
38  add-disk:
39      cat /sys/class/zram-control/hot_add
40
41  get-comp-algorithm:
42      cat /sys/block/zram$(id)/comp_algorithm
43
44  set-comp-algorithm:
45      echo $(algorithm) > /sys/block/zram$(id)/comp_algorithm
46
47  set-disksize:
48      echo $(disksize) > /sys/block/zram$(id)/disksize
49
50  memory-stat:
51      cat /sys/block/zram$(id)/mm_stat
52
53  reset-disk:
54      echo 1 > /sys/block/zram$(id)/reset
55
56  remove-disk:
57      echo $(id) > /sys/class/zram-control/hot_remove
```

3.4 Тестирование разработанного программного обеспечения

Тестирование разработанного программного обеспечения проводилось с помощью тестовых наборов, которые представляют собой архивы, созданные с помощью утилиты tar [34] следующей командой: `tar cf name.tar name`, где

Таблица 3.1 – Команды конфигурационного файла

Команда	Аргументы	Описание
make	-	собрать загружаемый модуль
make load	-	загрузить собранный модуль
make info	-	проверить, что модуль загружен
make log	-	получить сообщения системного журнала
make unload	-	выгрузить модуль
make add-disk	-	добавить диск, будет выведен его идентификатор
make get-comp-algorithm	id — идентификатор диска	получить список доступных алгоритмов сжатия
make set-comp-algorithm	algorithm — алгоритм сжатия id — идентификатор диска	установить алгоритм сжатия
make set-disksize	disksize — размер диска id — идентификатор диска	установить размер диска
make memory-stat	id — идентификатор диска	получить статистику памяти диска
make reset-disk	id — идентификатор диска	освободить память диска
make remove-disk	id — идентификатор диска	удалить диск

name — имя тестового набора.

В соответствии с выделенными классами эквивалентности были созданы следующие тестовые наборы:

- img-test.tar — архив, составленный из файлов с расширением .jpg из открытого набора изображений цветов [35];
- text-test.tar — архив, состоящий из текстовых файлов исходного кода ядра Linux [36].

Загрузка данных в устройство выполнялась следующей командой: `dd if=name.tar of=/dev/zram0`, где name — имя тестового набора.

При проведении тестирования диски zram на виртуальных машинах были настроены так:

- был выбран алгоритм сжатия zstd;
- размер устройства установлен в один гигабайт.

Пороговое значение для принятия решения о сжатии было установлено в 34 000.

В листинге 3.4 приведены результаты тестирования несжимаемых страниц памяти.

Листинг 3.4 — Результаты тестирования несжимаемых данных

```
1  Без оптимизации:
2
3  + исходный объем (Б):           8151040
4  + объем после сжатия (Б):       8071963
5  + число несжатых страниц:       1914
6  + число страниц с повторяющимся элементом: 2
7
8  С оптимизацией:
9
10 + исходный объем (Б):           8151040
11 + объем после сжатия (Б):       8071963
12 + число несжатых страниц:       1914
13 + число страниц с повторяющимся элементом: 2
```

В листинге 3.5 представлены результаты тестирования сжимаемых страниц памяти.

Листинг 3.5 — Результаты тестирования сжимаемых данных

```
1  Без оптимизации:
2
3  + исходный объем (Б):           8552448
4  + объем после сжатия (Б):       2811680
5  + число несжатых страниц:       0
6  + число страниц с повторяющимся элементом: 1
7
8  С оптимизацией:
9
10 + исходный объем (Б):           8552448
11 + объем после сжатия (Б):       2811680
12 + число несжатых страниц:       0
13 + число страниц с повторяющимся элементом: 1
```


Сравнение полученных результатов тестирования показало, что тесты пройдены успешно.

Вывод

В данном разделе был обоснован выбор языка программирования С и утилиты make в качестве средств программной реализации. Были представлены детали реализации программного решения. Кроме того были изложены команды для сборки и запуска программного обеспечения и настройки устройства zgam. Были описаны результаты тестирования разработанного программного обеспечения.

4 Исследовательский раздел

Вычисляемое значение информационной энтропии разработанным методом подсчета должно коррелировать с показателями качества сжатия, описанными в подразделе 1.2.2: коэффициентом сжатия, экономией пространства и силой сжатия. В соответствии с этим проводится исследование корреляции получаемого значения энтропии и указанных характеристик.

Так как на качество сжатия помимо алгоритма сжатия влияет их структура, производится исследование зависимости показателей качества сжатия и времени сжатия от различных типов хранимых в памяти данных.

Для анализа временной эффективности разработанной оптимизации выполняется исследование соотношения времени сжатия и времени вычисления энтропии при записи данных в устройство `zram`.

Для проведения исследований используются следующие данные:

- `exec-test.tar` — архив, содержащий исполняемые файлы из директории `/usr/bin`;
- `text-test.tar` — архив, состоящий из текстовых файлов исходного кода ядра Linux [36];
- `pdf-test.tar` — архив, включающий в себя файлы с расширением `.pdf` из открытого набора резюме `livecareer.com` [37];
- `img-test.tar` — архив, составленный из файлов с расширением `.jpg` из открытого набора изображений цветов [35].

4.1 Исследование корреляции энтропии и показателей качества сжатия

На рисунке 4.1 приведена матрица корреляции, составленная с помощью коэффициента корреляции Пирсона. Построенная матрица корреляции устанавливает зависимости между следующими характеристиками:

- значением энтропии, вычисляемой разработанным методом подсчета;
- значением энтропии, определяемой методом скользящего окна;

- значением энтропии, подсчитанной биномиальным методом;
- коэффициентом сжатия;
- экономией пространства;
- силой сжатия.

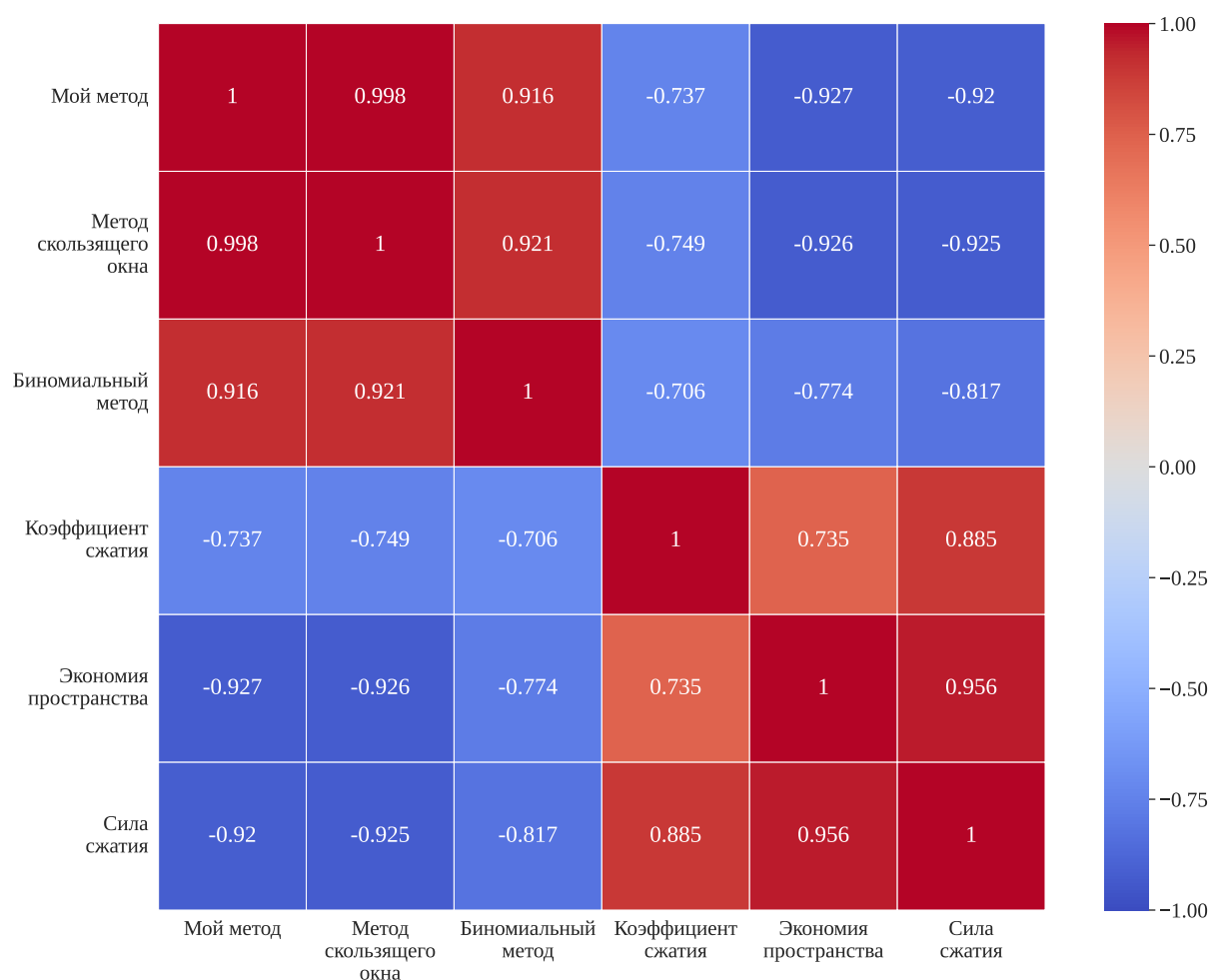


Рисунок 4.1 – IDEF0-диаграмма первого уровня

В результате анализа полученной матрицы корреляции можно сделать следующие выводы:

1. Значения энтропии, вычисляемые описанными методами подсчета, обладают высокой отрицательной корреляцией с показателями качества сжатия. При этом корреляция значений энтропии с показателями качества сжатия, определяемых разработанным методом и методом скользящего окна, выше корреляции энтропии, получаемой биномиальным методом.

2. Значения энтропии, вычисляемые указанными методами подсчета, обладают высокой положительной корреляцией друг с другом.

Так, энтропия, вычисляемая разработанным методом, обладает высокой отрицательной корреляцией с показателями качества сжатия, сравнимой с корреляцией энтропии, определяемой методом скользящего окна, и большей корреляции энтропии, получаемой биномиальным методом.

4.2 Исследование зависимости характеристик от типов хранимых в памяти данных

На рисунках 4.2-4.5 представлены зависимости коэффициента сжатия, экономии пространства, силы сжатия и времени сжатия от формата сжимаемых данных алгоритмами сжатия zstd, lz4, lz4hc, lzo и lzo-rle.

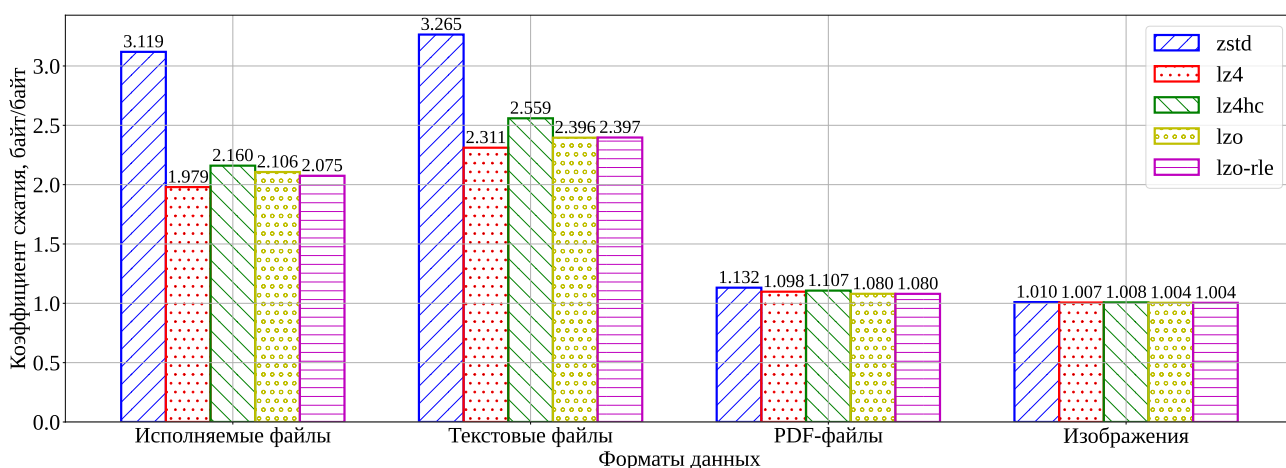


Рисунок 4.2 – Зависимость коэффициента сжатия от формата данных

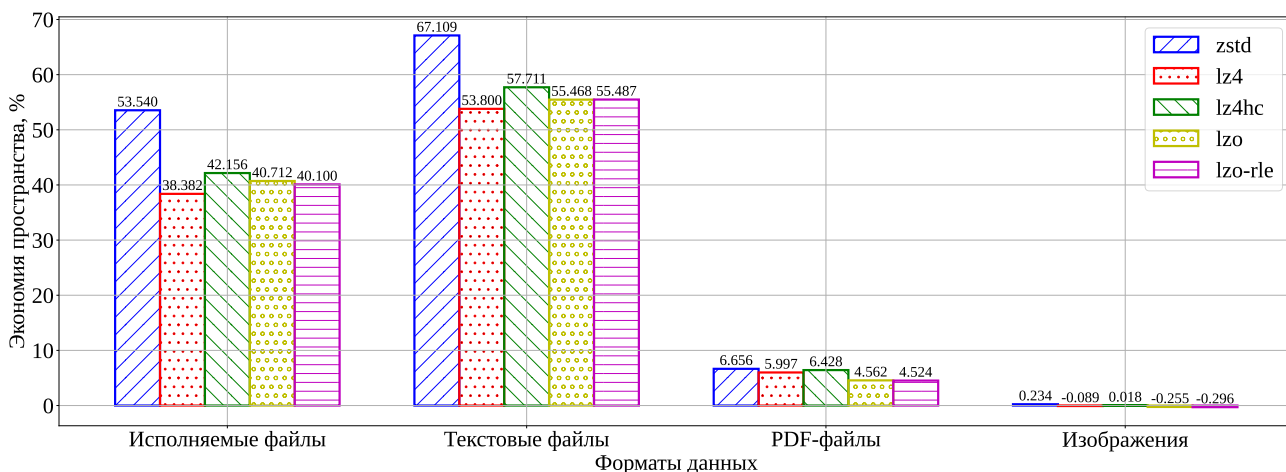


Рисунок 4.3 – Зависимость экономии пространства от формата данных

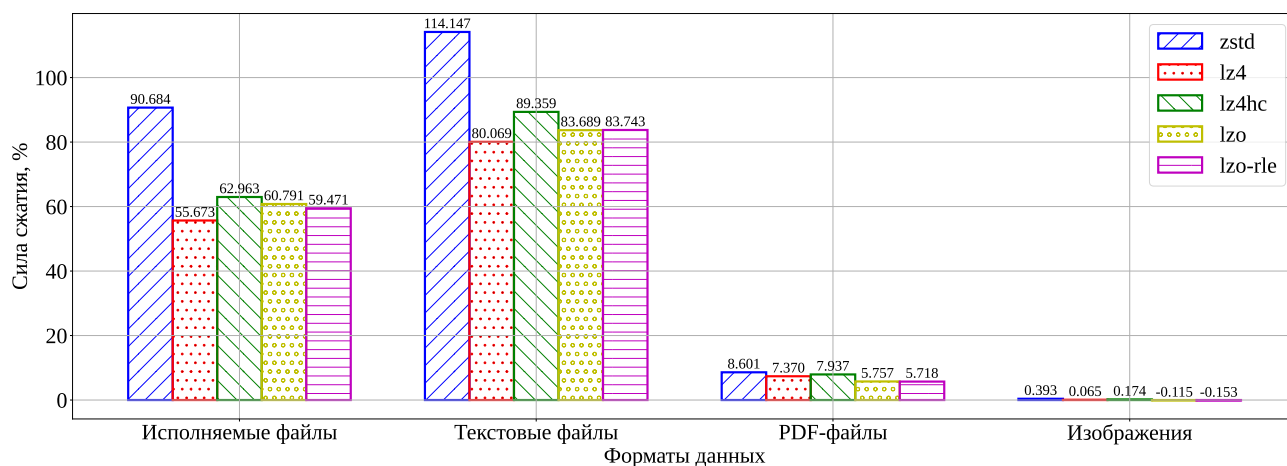


Рисунок 4.4 – Зависимость силы сжатия от формата данных

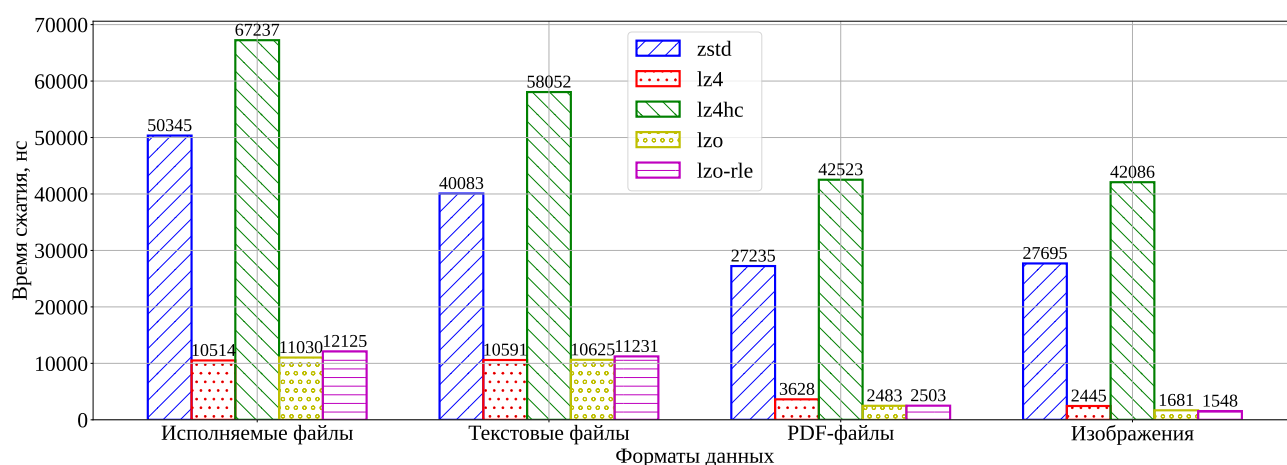


Рисунок 4.5 – Зависимость времени сжатия от формата данных

В результате анализа построенных гистограмм можно сделать следующие выводы:

1. Наибольшие значения показателей качества сжатия у текстовых данных.
2. Значения показателей качества сжатия у исполняемых файлов меньше, чем у текстовых, но больше, чем у PDF-файлов.
3. Наименьшие значения показателей качества сжатия у изображений.
4. Наибольшее время сжатия у исполняемых файлов.
5. Время сжатия текстовых данных меньше, чем у исполняемых, но больше, чем у PDF-файлов.
6. Наименьшее время сжатия у изображений.

Данные с расширениями .pdf и .jpg являются сжатыми данными, поэтому значения показателей качества для указанных данных меньше, чем для текстовых и исполняемых данных.

4.2.1 Исследование временной эффективности

В таблице 4.1 показано соотношение времени сжатия и времени вычисления информационной энтропии для доступных в модулей zgram алгоритмов сжатия: zstd, lz4, lz4hc, lzo и lzo-rle. Столбчатые диаграммы, соответствующие полученным результатам, приведены на рисунке 4.6.

Таблица 4.1 – Соотношение времени сжатия и времени вычисления энтропии

Алгоритм сжатия	Сжатие, %	Вычисление энтропии, %	Другие функции, %
zstd	84	12	4
lz4	58	32	10
lz4hc	89	8	3
lzo	61	31	8
lzo-rle	60	31	9

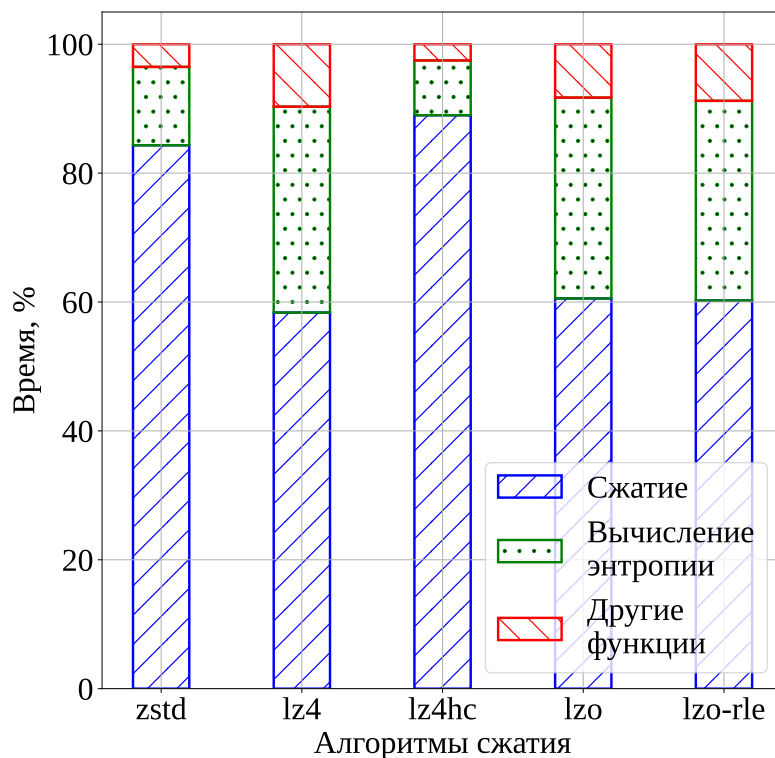


Рисунок 4.6 – Соотношение времени сжатия и времени вычисления энтропии

В результате анализа построенных гистограмм можно сделать следующие выводы:

1. Время вычисления энтропии в семь раз меньше времени сжатия страниц для алгоритма сжатия `zstd`.
2. Время вычисления энтропии почти в 1.81 раз меньше времени сжатия страниц для алгоритма сжатия `lz4`.
3. Время вычисления энтропии в 11.13 раз меньше времени сжатия страниц для алгоритма сжатия `lz4hc`.
4. Время вычисления энтропии в 1.97 раз меньше времени сжатия страниц для алгоритма сжатия `lzo`.
5. Время вычисления энтропии в 1.94 раз меньше времени сжатия страниц для алгоритма сжатия `lzo-rle`.

Таким образом, время вычисления энтропии значительно ниже времени сжатия данных.

Вывод

В данном разделе было проведено исследование корреляции показателей качества сжатия и значений информационной энтропии, вычисляемых разработанным методом, методом скользящего окна и биномиальным методом. Построенная с использованием коэффициента Пирсона матрица корреляции показала, что определяемое значение энтропии обладает высокой отрицательной корреляцией с показателями качества сжатия, поэтому может использоваться для оценки качества сжатия.

Была установлена зависимость показателей качества сжатия от типов хранимых в памяти данных: для PDF-файлов и jpg-файлов, которые являются сжатыми данными, характеристики сжатия ниже, чем их значения для текстовых и исполняемых данных.

В результате анализа временной эффективности было выявлено, что время вычисления энтропии значительно ниже времени сжатия. Таким образом, разработанная оптимизация позволяет увеличить скорость обработки несжимаемых страниц памяти в модуле сжатия `zgam` и производительность всей системы в целом.

ЗАКЛЮЧЕНИЕ

В результате выполнения данной работы была разработана оптимизация метода сжатия страниц памяти с использованием подсчета информационной энтропии.

В ходе выполнения работы были решены следующие задачи:

- рассмотрены основные понятия управления памятью в операционной системе;
- проведен анализ предметной области сжатия данных,
- изучены свойства информационной энтропии и ее связь со сжатием данных;
- описаны известные методы подсчета информационной энтропии;
- сформулированы основные этапы оптимизации метода сжатия страниц памяти с использованием подсчета информационной энтропии;
- разработано программное обеспечение, реализующее данную оптимизацию метода;
- протестировано разработанное программное обеспечение;
- исследована корреляция информационной энтропии и показателей качества сжатия;
- установлены зависимости времени сжатия и коэффициента сжатия от различных типов хранимых в памяти данных;
- исследована временная эффективность разработанной оптимизации.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Коржова Д. А., Смеричевский Э. Ф. Информационное общество: к анализу понятия // Вестник Науки и Творчества. — 2019. — С. 9—13.
2. Ranjan A., Raha A. Approximate Memory Compression // IEEE Transactions on Very Large Scale Integration (VLSI) Systems. — 2020.
3. Филиппов А. А. Операционные системы: учебное пособие // Ульяновск: Изд-во Ульян. тех. ун-та. — 2021. — С. 100.
4. Столлингс В. Операционные системы: внутренняя структура и принципы проектирования // СПб.: ООО «Диалектика». — 2020. — С. 1264.
5. Silberschatz A., Galvin P. B., Gagne G. Operating System Concepts // New York: Wiley. — 2018. — С. 1278.
6. Uthayakumar J., Vengattaraman T., Dhavachelvan P. A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications // Journal of King Saud University — Computer and Information Sciences. — 2021. — С. 119—140.
7. Пантелеев Е. Р., Алыкова А. Л. Алгоритмы сжатия данных без потерь: учебное пособие для вузов // СПб.: Лань. — 2022. — С. 172. с.
8. Ranjan A., Raha A., Raghunathan V. L2C: Combining Lossy and Lossless Compression on Memory and I/O // ACM Trans. Embed. Comput. Syst. — 2022. — С. 1-27.
9. Данные, не поддерживающие сжатие [Электронный ресурс]. — Режим доступа URL: https://www.ibm.com/docs/ru/informix-servers/12.10?topic=data_that_you_cannot_compress (Дата обращения: 07.05.2023).
10. Березкин Е. Ф. Основы теории информации и кодирования: учебное пособие // СПб.: Лань. — 2022. — С. 320.
11. Rodrigues M. Information-Theoretic Methods in Data Science // Cambridge: Cambridge University Press. — 2021. — С. 43.
12. Попов И. Ю., Блинова И. В. Теория информации // СПб.: Лань. — 2022. — С. 160.

13. *Осокин А. Н., Мальчуков А. Н.* Теория информации: учебное пособие для вузов // М.: Издательство Юрайт. — 2022. — С. 205.
14. *Zbili M., Rama S.* A Quick and Easy Way to Estimate Entropy and Mutual Information for Neuroscience // Frontiers in Neuroinformatics. — 2021.
15. *Cheng X., Li Z.* How does Shannon’s source coding theorem fare in prediction of image compression ratio with current algorithms? // International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. — 2020. — С. 1313—1319.
16. *Ryabko B.* Time-Universal Data Compression // Algorithms. — 2019.
17. *Guo H.* File Entropy Signal Analysis Combined With Wavelet Decomposition for Malware Classification // IEEE Access. — 2020. — С. 158961—158971.
18. *Пухальский Г. И., Новосельцева Т. Я.* Проектирование цифровых устройств: учебное пособие для вузов // СПб.: Лань. — 2022. — С. 896.
19. *Borysenko O.* On the binomial method for calculation of entropy // Grail of Science. — 2022. — С. 113—118.
20. *Рябко Б. Я., Фионов А. Н.* Криптография в информационном мире // М.: Горячая линия-Телеком. — 2018. — С. 300.
21. Что представляют инструменты с открытым исходным кодом? [Электронный ресурс]. — Режим доступа URL: <https://aws.amazon.com/ru/what-is/open-source/> (Дата обращения: 07.05.2023).
22. Desktop Operating System Market Share Worldwide [Электронный ресурс]. — Режим доступа URL: <https://gs.statcounter.com/os-market-share/desktop/worldwide> (Дата обращения: 07.05.2023).
23. The Linux Kernel Module Programming Guide [Электронный ресурс]. — Режим доступа URL: <https://sysprog21.github.io/lkmpg/#introduction> (Дата обращения: 07.05.2023).
24. *Fox R.* Linux with Operating System Concepts // Boca Raton: CRC Press. — 2022. — С. 100.
25. zram: Compressed RAM-based block devices [Электронный ресурс]. — Режим доступа URL: <https://docs.kernel.org/admin-guide/blockdev/zram.html?highlight=zram> (Дата обращения: 07.05.2023).

26. Kernel Crypto API Interface Specification [Электронный ресурс]. — Режим доступа URL: <https://docs.kernel.org/crypto/intro.html> (Дата обращения: 07.05.2023).
27. zswap [Электронный ресурс]. — Режим доступа URL: <https://docs.kernel.org/admin-guide/mm/zswap.html?highlight=zswap> (Дата обращения: 07.05.2023).
28. Лав Р. Ядро Linux: описание процесса разработки // М.: ООО «И.Д. Вильямс». — 2013. — С. 496.
29. zram - drivers/block/zram - Linux source code (v5.15.71) - Bootlin [Электронный ресурс]. — Режим доступа URL: <https://elixir.bootlin.com/linux/v5.15.71/source/drivers/block/zram> (Дата обращения: 07.05.2023).
30. Why don't some files compress very much? [Электронный ресурс]. — Режим доступа URL: <https://kb.corel.com/en/125893> (Дата обращения: 07.05.2023).
31. Данные, поддерживающие сжатие [Электронный ресурс]. — Режим доступа URL: https://www.ibm.com/docs/ru/informix-servers/12.10?topic=data_that_you_can_compress (Дата обращения: 07.05.2023).
32. The GNU C Reference Manual [Электронный ресурс]. — Режим доступа URL: <https://www.gnu.org/software/gnu-c-manual/gnu-c-manual.html> (Дата обращения: 07.05.2023).
33. GNU make [Электронный ресурс]. — Режим доступа URL: <https://www.gnu.org/software/make/manual/make.html#Simple-Makefile> (Дата обращения: 07.05.2023).
34. GNU Tar [Электронный ресурс]. — Режим доступа URL: <https://www.gnu.org/software/tar/> (Дата обращения: 07.05.2023).
35. 102 Category Flower Dataset [Электронный ресурс]. — Режим доступа URL: <https://www.robots.ox.ac.uk/~vgg/data/flowers/102/index.html> (Дата обращения: 07.05.2023).
36. The Linux Kernel Archives [Электронный ресурс]. — Режим доступа URL: <https://kernel.org/> (Дата обращения: 07.05.2023).

37. Resume Dataset [Электронный ресурс]. — Режим доступа URL: <https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset?resource=download> (Дата обращения: 07.05.2023).

ПРИЛОЖЕНИЕ А