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»

Лабораторная работа № 5
по дисциплине «Анализ Алгоритмов»

Тема Организация параллельных вычислений по конвейерному принципу

Студент Шахнович Дмитрий Сергеевич

Группа ИУ7-52Б

Преподаватели Волкова Л.Л., Строганов Д.В.

Москва, 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Параллелизм описывает последовательности, которые происходят одновременно [1]. Таким образом параллельные вычисления таковы, что они выполняются одновременно. Распараллеливание вычислений может привести к росту временной эффективности программы при использовании на многопроцессорных и однопроцессорных, если программа часто блокируется ожиданиями ввода/вывода, системах [2].

Некоторые программы можно разделить на части, при этом каждая такая часть выполняет определённую работу над данными и передаёт их следующей части. Таким образом образуется конвейер по обработке данных, при этом части этого конвейера могут работать параллельно друг другу.

Целью данной работы является разработка ПО, выполняющего скачивание страниц и парсинг рецептов с сайта menunedeli.ru с помощью параллельных вычислений по конвейерному принципу.

Задачами работы являются:

— рассмотрение структуры сайта;

— разработка ПО, выполняющего парсинг рецептов в конвейере;

— исследование характеристик разработанного ПО на данных о интервалах времени стадий обработки рецепта.

1 Входные и выходные данные

Входными данными для программы являются ссылки на страницы сайта menunedeli.ru. с рецептами. Каждая ссылка содержит один рецепт в веб-ресурсе. Выходными данными являются файлы json – загруженные рецепты по ссылкам из входных данных, при этом в каждом json файле хранится:

— название рецепта;

— ссылка на изображение блюда;

— список ингредиентов;

— шаги рецепта.

2 Преобразование входных данных в выходные

Для получения ссылок с рецептами с сайта в виде файла разработан скрипт parse.py на языке python3 [3]:

Листинг 2.1 — parse.py – скрипт для получения ссылок с рецептами

```
import requests as re
import bs4
```

```
import argparse
import sys

parser = argparse.ArgumentParser(
prog="parse",
description="Скачивает ссылки на статьи с сайта menunedeli.ru"
)
parser.add_argument("-c", "--count", type=int, default=1000, help="Колич
    ество ссылок для скачивания")
parser.add_argument("-s", "--save", type=str, default="links.txt", help=
    "Файл, куда сохранять ссылки")

args = parser.parse_args()

saveTo = args.save
UpToLinks = args.count
catalogFormat = "https://menunedeli.ru/novye-stati/page/{}"
with open(saveTo, "w") as f:
  links = 0
  page = 1
  while True:
    catalogPage = re.get(catalogFormat.format(page))

    bs = bs4.BeautifulSoup(catalogPage.content, "lxml")

    for a in bs.find_all('article'):
      for link in a.find_all('meta', attrs={'itemprop':'url'}):
        if link['content'].startswith("https://menunedeli.ru/recipe"):
          print(link['content'], file=f)
          links += 1
          if links >= UpToLinks:
            exit(0)
    page += 1
```

Полученный в результате работы программы 2.1 файл выступает входными данными для разработанного ПО.

Основная часть программы написана на языке GO [4] использую паттерн проектирования конвейер (pipeline) [5]. Конвейер состоит из 3-х частей:

1) скачивание страницы;

2) вычленение из страницы частей рецепта (парсинг);

3) сохранение рецепта в базе данных.

Каждая из частей запускается в отдельной горутине (goroutine) [6], которые передают данные через каналы (channel) [6]

Элементом данных выступает структура, которая хранит в себе информацию о рецепте, информацию о моментах времени начала и конца обработки отдельной частью программы и метаинформацию для обработчиков конвейера. Структура представлена в листинге 2.2.

Листинг 2.2 — Данные, передающиеся между частями конвейера

```
// Передаётся между частями конвейера
type PipelineUnit struct {
  recipedb.Recipe
  PipelineTimes
  PageReader io.ReadCloser
  Err        error
}

type Ingredient struct {
  Name   string
  Amount string
  Unit   string
}

type Recipe struct {
  ID          int64
  IssueID     int64
  Url         string
  Title       string
  ImageURL    string
  Ingredients []Ingredient
  Steps       []string
}

type PipelineTimes struct {
  LoadingStart time.Time
  LoadingEnd   time.Time
  ParsingStart time.Time
  ParsingEnd   time.Time
  StorageStart time.Time
  StorageEnd   time.Time
}
```

Полученные в результате работы программы рецепты сохраняются в базе данных PostgreSQL [7], а из неё отдельным скриптом сохраняются в виде json-файла.

3 Примеры работы программы

Один из примеров обработки веб-страниц [8] представлен на рисунке 3.1 и листинге 3.1.

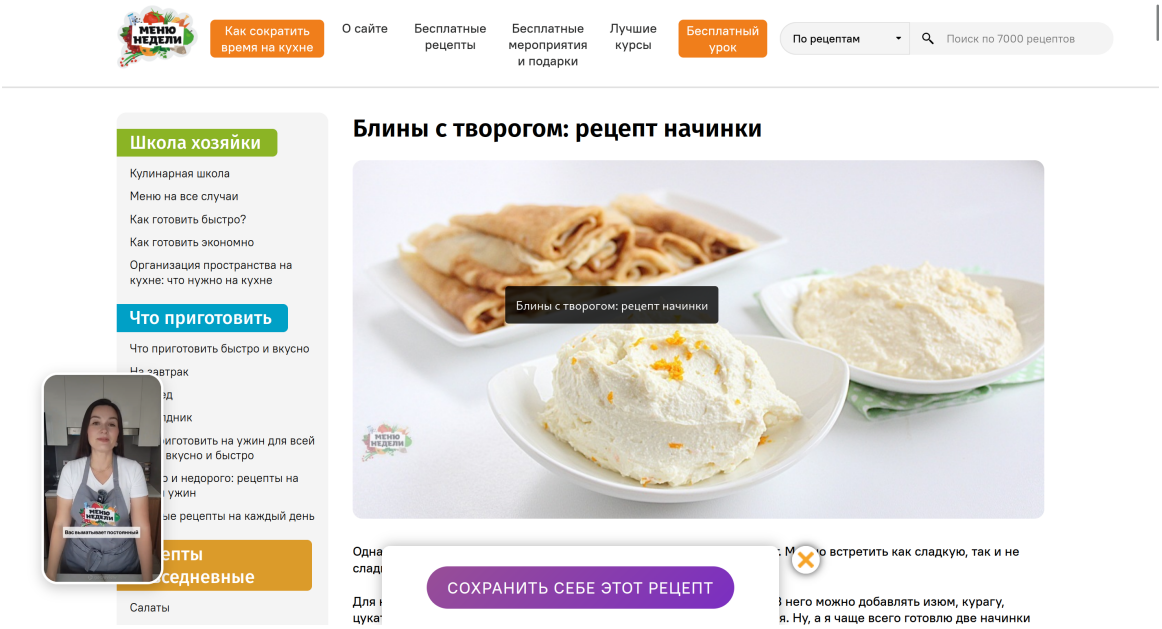

Рисунок 3.1 — Веб-страница по ссылке [8]

Листинг 3.1 — json рецепта после обработки


```
{
  "ID": 528,
  "ImageURL": "https://menunedeli.ru/wp-content/uploads/2024/03/Bliny-s-
      tvorogom-recept-nachinki-500x350.jpg-500x333.jpg?v=1710428422",
  "Ingredients": [
  {
    "Amount": "500",
    "Name": "Творог",
    "Unit": "г"
  },
  {
    "Amount": "80",
    "Name": "Сахар",
    "Unit": "г"
  },
  {
    "Amount": "1",
    "Name": "Яйцо куриное",
    "Unit": "шт."
  },
  {
    "Amount": "2",
```

```
      "Name": "Сметана",
      "Unit": "ст.л."
    },
    {
      "Amount": "0.5",
      "Name": "Цедра апельсина",
      "Unit": "ч.л."
    },
    {
      "Amount": "",
      "Name": "Ваниль",
      "Unit": "по вкусу"
    }
    ],
    "IssueID": 9161,
    "Steps": [
    "...",
    "...",
    ...
    ],
    "Title": "Блины с творогом: рецепт начинки",
    "Url": "https://menunedeli.ru/recipe/bliny-s-tvorogom-recept-nachinki/
       "
  }
```


4 Тестирование

Тестирование программы проводилось загрузкой в качестве входных данных 1000 ссылок на различные рецепты веб-ресурса menunedeli.ru. При этом отслеживалось количество удачных скачиваний и обработок страницы, а также выборочная ручная проверка 10 случайных выходных файлов.

Тестирование было успешно пройдено.

5 Описание исследования

Технические характеристики устройства,на котором выполнялась программа:

— процессор: AMD Ryzen 7 5800H (16) @ 4.46 ГГц;

— оперативная память: 16 ГБ;

— операционная система: Arch Linux x86_64.

На вход программе был подан файл с 1000 ссылок на рецепты сайта menunedeli.ru, каждая из которых была успешно обработана.

Во время выполнения, в программе фиксируются моменты времени начала и конца обработки рецепта очередным этапом конвейера. Перед окончанием программы эти данные сохраняются в отдельный файл. Часть этого файла представлена в листинге 5.1.

Листинг 5.1 — Часть лога, сформированного конвейром

```
3 loading 0.184901  0.209007
2 storage 0.205416  0.227645
4 loading 0.209011  0.246545
3 parsing 0.209074  0.228591
3 storage 0.228596  0.237616
5 loading 0.246548  0.281826
4 parsing 0.246586  0.263061
4 storage 0.263066  0.282869
```

Где 1-й столбец – идентификатор заявки, 2-ой – тип обработчика, 3-й – начало обработки, 4-й – конец обработки.

Из лога 5.1 можно увидеть, что в случае конвейерной обработки может быть простой отдельных её частей, например после окончания обработки заявки 3, storage простаивает до окончания обработки 4-ой заявки обработчиком parsing. Соответственно storage простаивает между 0.237616 и 0.263066.

Простои обработчиков и части лога 5.1 можно увидеть на рисунке 5.1

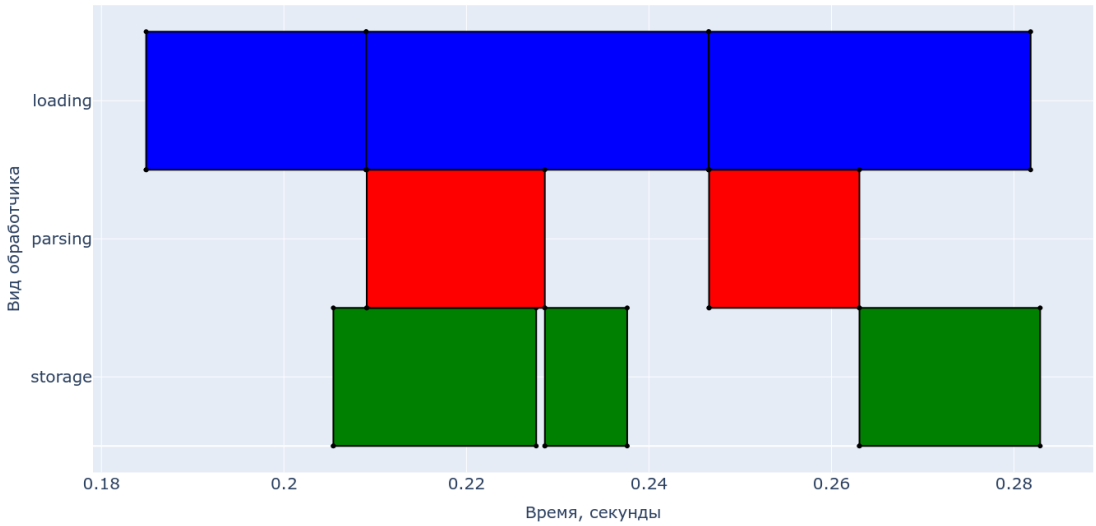

Рисунок 5.1 — Временная диаграмма простоя обработчиков

На данном участке обработчик загрузки страниц беспрерывно их скачивают, а обработчики вычленения и сохранения рецепта простаивают в ожидании новых данных.

Также по данным всего лога программы была составлена таблица с временными характеристиками программы 5.1.

Таблица 5.1 — Временные характеристики разработанной программы

Характеристика	Минимум, мс	Максимум, мс	Среднее, мс	Медиана, мс
Время загрузчики	16.09	610.20	33.04	22.30
Время парсинг	12.31	41.62	15.70	15.05
Время сохранения	7.29	50.72	17.40	16.98
Время ожидания парсера	0.002	22.27	0.31	0.03
Время ожидания сохранения	0.02	30.99	0.70	0.06
Время простоя парсера	0.02	596.90	17.53	7.24
Время простоя сохранения	0.02	614.60	15.83	5.27
Время обработки ссылки	40.5	663.6	67.16	56.16

В результате исследования сделан вывод, что в случае выполнения конвейерных параллельных вычислений могут возникать простои отдельных её частей, то есть моменты времени, когда часть конвейера не выполняет вычисления, ожидая данных от предыдущих частей конвейера.

ЗАКЛЮЧЕНИЕ

Целью – разработка ПО, выполняющего скачивание страниц и парсинг рецептов с сайта menunedeli.ru с помощью параллельных вычислений по конвейерному принципу – была выполнена.

В ходе работы были решены следующие задачи:

— рассмотрена структуры сайта;

— разработано ПО, выполняющее парсинг рецептов в конвейере;

— исследованы характеристики разработанного ПО на данных о интервалах времени стадий обработки рецепта.

В ходе исследования было выявлено, что использование конвейера для параллельных вычислений может привести к простою отдельных его частей из-за ожидания завершения обработки данных предыдущими частями конвейера.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. David R. Butenhof Programming with POSIX threads [Текст] / David R. Butenhof — . — : Library of Congress Cataloging-in-Publication Data, 1997 — 202 с.

2. Таненбаум Э., Бос Х. Современные операционные системы. [Текст] / Таненбаум Э., Бос Х. — 4-е изд.. — СПб.: Питер, 2015 — 1120 с.

3. Python / [Электронный ресурс] // Python : [сайт]. — URL: https://www.python.org/ (дата обращения: 20.09.2024).

4. GO // GO programming language URL: https://go.dev/ (дата обращения: 24.10.2024).

5. Mario Castro Contreras. Go Design Patterns. - Birmingham: Packt Publishing Ltd., 2017. - 377 с.

6. The Go Memory Model // GO programming language URL: https://go.dev/ref/mem (дата обращения: 24.10.2024).

7. PostgreSQL: The World's Most Advanced Open Source Relational Database // PostgreSQL URL: https://www.postgresql.org/ (дата обращения: 24.10.2024).

8. Ольга Котельникова. Блины с творогом: рецепт начинки. / Ольга Котельникова. [Электронный ресурс] // Меню недели : [сайт]. — URL: https://menunedeli.ru/recipe/bliny-s-tvorogom-recept-nachinki/ (дата обращения: 20.10.2024).