PARALLEL AND DISTRIBUTED COMPUTING

ASIIGNMENT NO. 2

Parallelized Web Crawler with Load Balancing

Group Members:

NAME	CMS ID
Muhammad Ammar bin Akram	414563
Malik Muhammad Aman	409918
Hannan Yousaf Butt	405326

1. Introduction

This report analyzes the performance of a web crawler implemented in three versions:

- **Serial (Single-threaded)**
- **Multithreaded**
- **MPI-based Distributed**

The goal was to evaluate the **execution time**, understand the **design trade-offs**, and assess the **scalability** of each version using a fixed URL workload.

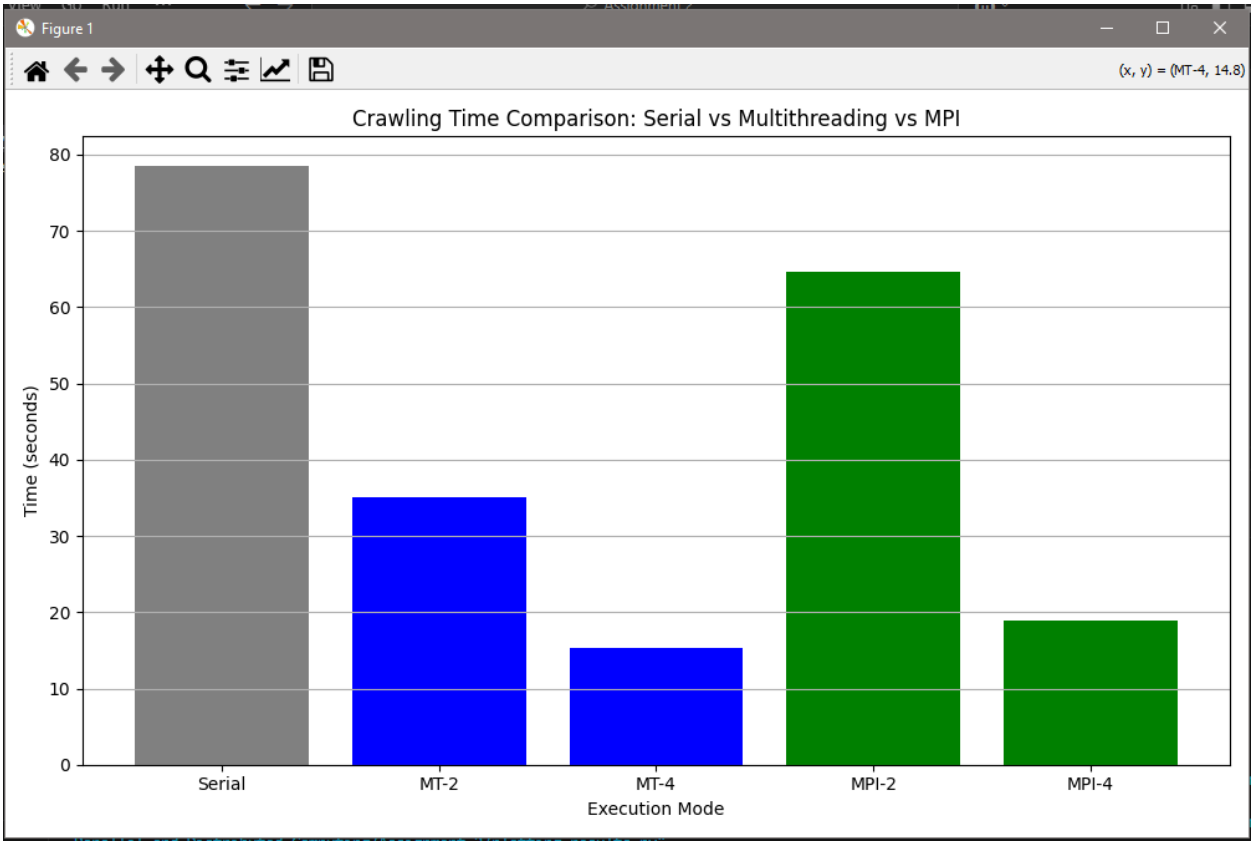
2. Benchmark Results

Below are the crawl timings obtained from crawling a fixed set of web pages (50):

Version	Configuration	Time taken
Sequential	-	78.47 seconds
Multithreaded	2 threads	35.01 seconds
Multithreaded	4 threads	15.31 seconds
MPI (Distributed)	2 workers	64.68 seconds
MPI (Distributed)	4 workers	18.85 seconds

Visual Comparison:

A bar chart was generated to visually compare the performance. It shows that both multithreading and MPI significantly reduce the crawl time, especially when the number of workers/threads increases.



3. Design Trade-offs

Multithreading:

- **Pros:**
  - Easy to implement and deploy
  - Shared memory leads to low communication overhead
  - Scales well up to moderate thread counts
- **Cons:**
  - Python's Global Interpreter Lock (GIL) can limit CPU-bound performance
  - Threads share memory — potential for data corruption without locks

MPI-based Distributed Crawling:

- **Pros:**
  - Suitable for large-scale distributed environments
  - Can scale across multiple physical machines
- **Cons:**
  - Higher communication overhead due to inter-process messaging
  - Requires MPI setup, making deployment more complex
  - Debugging across processes is harder than with threads

4. Scalability Analysis

Speedup (vs. Serial):

- Multithreaded (2 threads): 78.47 / 35.01 ≈ **2.24×**
- Multithreaded (4 threads): 78.47 / 15.31 ≈ **5.12×**
- MPI (2 workers): 78.47 / 64.68 ≈ **1.21×**
- MPI (4 workers): 78.47 / 18.85 ≈ **4.16×**

Observations:

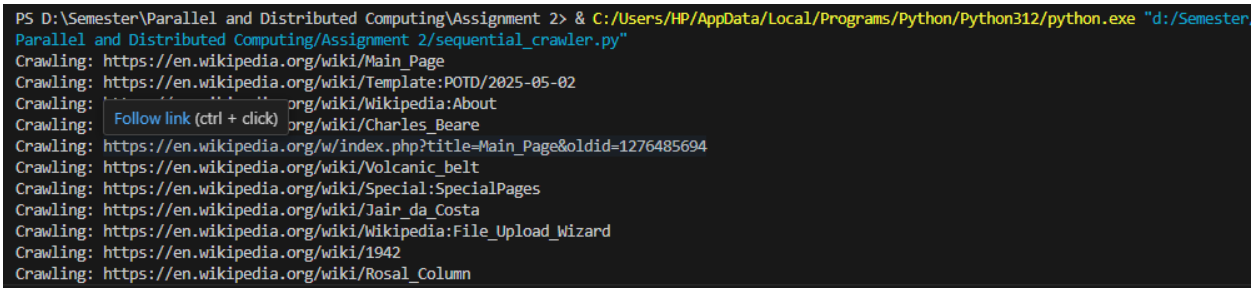
- **Multithreading** achieves better scaling at low thread counts due to minimal overhead.
- **MPI** shows improved performance with more workers, but suffers from higher overhead at low counts (2 workers).
- **Diminishing returns** begin to appear as the number of threads/workers increases, especially if the workload is not large enough to fully utilize them.

5. Conclusion & Recommendations

- For **moderate workloads**, **multithreading is the most efficient** in terms of simplicity and performance.
- For **scalable distributed crawling across multiple machines**, **MPI is more suitable**, albeit with increased complexity.
- To further improve performance:
    - Optimize HTML parsing (e.g., limit tags extracted)
    - Avoid re-crawling duplicate URLs
    - Consider asynchronous I/O (e.g., using `aiohttp`) for I/O-bound performance

ScreenShots of Code Running:

Sequential Crawler:



```
Crawled 50 pages in 88.36 seconds

Crawl Results:
1. Wikipedia, the free encyclopedia - https://en.wikipedia.org/wiki/Main_Page
2. Template:POTD/2025-05-02 - Wikipedia - https://en.wikipedia.org/wiki/Template:POTD/2025-05-02
3. Wikipedia:About - Wikipedia - https://en.wikipedia.org/wiki/Wikipedia:About
4. Charles Beare - Wikipedia - https://en.wikipedia.org/wiki/Charles_Beare
5. Wikipedia, the free encyclopedia - https://en.wikipedia.org/w/index.php?title=Main_Page&oldid=1276485694
6. Volcanic belt - Wikipedia - https://en.wikipedia.org/wiki/Volcanic_belt
7. Special pages - Wikipedia - https://en.wikipedia.org/wiki/Special:SpecialPages
8. Jair da Costa - Wikipedia - https://en.wikipedia.org/wiki/Jair_da_Costa
9. Wikipedia:File upload wizard - Wikipedia - https://en.wikipedia.org/wiki/Wikipedia:File_Upload_Wizard
10. 1942 - Wikipedia - https://en.wikipedia.org/wiki/1942
```

Multi-Threaded Crawler:

```
PS D:\Semester\Parallel and Distributed Computing\Assignment 2> & C:/Users/HP/AppData/Local/Programs/Python/Python312/python.exe "d:/Semester/
Parallel and Distributed Computing/Assignment 2/task2-optionA.py"
Using 2 threads for crawling
Crawling Progress:   0%|                                                                                | 0/50 [00:00<?, ?it/s][
Worker-1] Crawled: https://www.youtube.com - Title: YouTube
Crawling Progress:   4%|                                                                              | 2/50 [00:00<00:17,  2.75it/s][
Worker-1] Crawled: https://www.youtube.com/ - Title: YouTube
Crawling Progress:   8%|                                                                              | 4/50 [00:01<00:20,  2.23it/s][
Worker-2] Crawled: https://www.youtube.com/about/ - Title: About YouTube - YouTube
Crawling Progress:  10%|                                                                              | 5/50 [00:02<00:23,  1.95it/s][
Worker-2] Crawled: https://www.youtube.com/about/copyright/ - Title: YouTube Copyright Rules & Policies - How YouTube Works
Crawling Progress:  12%|                                                                              | 6/50 [00:03<00:25,  1.71it/s][
Worker-1] Crawled: https://www.youtube.com/about/press/ - Title: Official YouTube Blog for Latest YouTube News & Insights
```

```
PS D:\Semester\Parallel and Distributed Computing\Assignment 2> & C:/Users/HP/AppData/Local/Programs/Python/Python312/python.exe "d:/Semester/
Parallel and Distributed Computing/Assignment 2/task2-optionA.py"
Using 4 threads for crawling
Crawling Progress:   0%|                                                                                | 0/50 [00:00<?, ?it/s][
Worker-1] Crawled: https://www.youtube.com - Title: YouTube
Crawling Progress:   4%|                                                                              | 2/50 [00:01<00:26,  1.83it/s][
Worker-1] Crawled: https://www.youtube.com/ - Title: YouTube
Crawling Progress:  12%|                                                                              | 6/50 [00:02<00:16,  2.72it/s][
Worker-4] Crawled: https://www.youtube.com/about/ - Title: About YouTube - YouTube
Crawling Progress:  14%|                                                                              | 7/50 [00:02<00:16,  2.57it/s][
Worker-1] Crawled: https://www.youtube.com/t/contact_us/ - Title: ہم سے رابطہ کریں
[Worker-3] Crawled: https://www.youtube.com/about/copyright/ - Title: YouTube Copyright Rules & Policies - How YouTube Works
Crawling Progress:  18%|                                                                              | 9/50 [00:03<00:12,  3.29it/s][
```

```
[Worker-1] Crawled: https://www.youtube.com/howyoutubeworks/our-commitments/fostering-child-safety/ - Title: How YouTube Keeps Kids Safe Onlin
e - How YouTube Works
[Worker-3] Crawled: https://www.youtube.com/howyoutubeworks/our-commitments/promoting-digital-wellbeing/ - Title: How Youtube Technology Suppo
rt Wellbeing - How YouTube Works
[Worker-2] Crawled: https://www.youtube.com/howyoutubeworks/our-commitments/sharing-revenue/ - Title: How YouTube Makes Money - How YouTube Wo
rks
Crawling Progress: 100%|                                                                              | 50/50 [00:18<00:00,  2.67it/s]

 Crawled 50 pages in 18.72 seconds.
 Avg speed: 2.67 pages/sec
Total pages crawled: 50
Time taken: 18.72 seconds
```

MPI based Crawler:

```
PS D:\Semester\Parallel and Distributed Computing\Assignment 2> mpiexec -n 3 python task2-optionB.py
Worker 2:
  URL: https://www.coursera.org
  Title: Coursera | Degrees, Certificates, & Free Online Courses
  Links found: 73
[Master] MPI Version: (2, 0)
Worker 1:
  URL: https://www.youtube.com
  Title: YouTube
  Links found: 8
[Master] Sending URL to worker 2: https://www.wikipedia.org
Worker 2:
  URL: https://www.wikipedia.org
```

```
PS D:\Semester\Parallel and Distributed Computing\Assignment 2> mpiexec -n 5 python task2-optionB.py
Worker 3:
  URL: https://www.wikipedia.org
  Title: Wikipedia
  Links found: 8
[Master] MPI Version: (2, 0)
Worker 1:
  URL: https://www.youtube.com
  Title: YouTube
  Links found: 8
[Master] Sending URL to worker 3: https://www.python.org
Worker 3:
  URL: https://www.python.org
```

```
  URL: http://planetpython.org/
  Title: Planet Python
  Links found: 857
[Master] Sending URL to worker 2: http://www.pylonsproject.org/
Worker 1:
  URL: https://pythoninsider.blogspot.com/2025/04/python-3140a7-3133-31210-31112-31017.html
  Title: Python Insider: Python 3.14.0a7, 3.13.3, 3.12.10, 3.11.12, 3.10.17 and 3.9.22 are now available
  Links found: 271
[Master] Sending URL to worker 3: http://bottlepy.org
Total crawl time: 16.21 seconds
Pages crawled by each worker: [15, 15, 12, 8]
Total pages crawled: 50
```

GitHub Repository Link:

https://github.com/Ammar-bin-Akram/Web-Crawlers