

OBAH ISAAC EBUKA
MMU ID: 23659765

HIGH PERFORMANCE COMPUTING AND BIG DATA
DEPARTMENT OF COMPUTING AND ENGINEERING.
THE MANCHESTER METROPOLITAN UNIVERSITY.
MANCHESTER, UNITED KINGDOM.

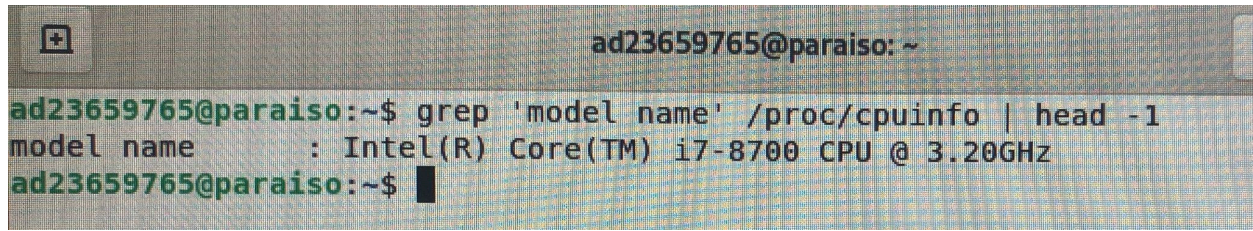
DECLARATION: I hereby affirm that this work is entirely my own and that all sources have been appropriately cited.

HIGH PERFORMANCE COMPUTING TASK

Before parallel computing, computers processed tasks sequentially, one at a time from start to finish. This method worked but was slow for large data problems because it couldn't use multiple processors simultaneously. With parallel computing, multiple tasks can be processed at the same time using multiple processors, greatly improving speed and efficiency (Eager et al., 1989). However, parallel computing has its own challenges, like making sure tasks are well-coordinated and managing communication between processors (Czamal et al., 2020).

In this report, I have analysed three versions of a Molecular Dynamics (MD) simulation program from the “~ad55142816/hpc_labs/assess.tar” folder which contains a serial version, an OpenMP, and an MPI parallelized version. The performance of each implementation is assessed, and potential improvements are discussed as well.

SYSTEM INFORMATION

A terminal window screenshot showing the command 'grep 'model name' /proc/cpuinfo | head -1' being executed. The output is 'model name : Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz'. The terminal title bar shows 'ad23659765@paraíso: ~'.

```
ad23659765@paraíso: ~  
ad23659765@paraíso:~$ grep 'model name' /proc/cpuinfo | head -1  
model name      : Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz  
ad23659765@paraíso:~$
```

IMPLEMENTATION OF THE SERIAL CODE

Commands for compilation of the serial code

Here are the steps carried out to compile the serial code

1. Navigate to the directory “HPC” where the “~ad55142816/hpc_labs/assess.tar” folder is located using the change directory “cd” command, Hence “cd HPC”.
2. Extract the files within the “~ad55142816/hpc_labs/assess.tar” folder using the command “tar -xf ~ad55142816/hpc_labs/assess.tar” which releases the files “md.c”, “md-MPI.c” and “md-OpenMP.c”.
3. Run the “ls” command to list the whole files in the directory to be sure that our serial code is available in the directory.
4. The serial code is within the “md.c” file. To view the content of this file, the command “gedit md.c” is run.
5. To compile the code, we use “gcc -O0 md.c -lm -o md_serial.exe”. This code uses the gcc compiler with no optimization to compile the md.c file into an executable md_serial.exe
6. Subsequent versions of the previous code were compiled but this time changing the optimization to “O1”, “O2” and “O3” and the respective time of compilation was recorded.

7. Steps **5 & 6** were done three consecutive times and the average was taken as the time for the compilation and a table of time “**T1**”, “**T2**”, and “**T3**” were drawn with their corresponding average “**average_time**” was calculated.
8. The final “**centre of mass**” was recorded as well.

IMPLEMENTATION OF THE MPI and OPENMP CODE

Commands for compilation of the MPI code

Here are the steps carried out to compile the MPI code

1. Navigate to the directory “**HPC**” where the “**~ad55142816/hpc_labs/assess.tar**” folder is located using the change directory “**cd**” command, Hence “**cd HPC**”.
2. Extract the files within the “**~ad55142816/hpc_labs/assess.tar**” folder using the command “**tar -xf ~ad55142816/hpc_labs/assess.tar**” which releases the files “**md.c**”, “**md-MPI.c**” and “**md-OpenMP.c**”.
3. Run the “**ls**” command to list the whole files in the directory to be sure that our serial code is available in the directory.
4. The MPI code is within the “**md-MPI.c**” file. To view the content of this file, the command “**gedit md-MPI.c**” is run. For OpenMP The OPENMP code is within the “**md-OpenMP.c**” file. To view the content of this file, the command “**gedit md-OpenMP.c**” is run.
5. To compile the code, we use “**mpicc -O0 md-MPI.c -lm -o md_mpi.exe**”. To compile MPI code, we use “**mpicc**” rather than “**gcc, cc or pgcc**”. This code uses the **mpicc** compiler with no optimization to compile the **md-MPI.c** file into an executable **md_mpi.exe**. For OPENMP use “**gcc -fopenmp -O0 md-OPENMP.c -lm -o md_open.exe**”. This code uses the **gcc** compiler with no optimization to compile the **md-OPENMP.c** file into an executable **md_openmp.exe**.
6. For hyperthreading **mpirun --use-hwthread-cpus -np 7 ./md_mpi.exe**
7. Run the executable using “**./md_mpi.exe**”. For **openmp** use “**./md_openmp.exe**”
8. Subsequent versions of the previous code were compiled but this time changing the optimization flag “**O0**” to “**O1**” - “**12**” at each instance and the respective time of compilation was recorded.
9. Steps **5 & 6** were done three consecutive times and the average was taken as the time for the compilation and a table of time “**T1**”, “**T2**”, and “**T3**” were drawn with their corresponding average “**average_time**” was calculated.
10. The final “**centre of mass**” was recorded as well.

TABLE OF TIME

MPI					OPENMP				
threads	T1	T2	T3	Average_time	threads	T1	T2	T3	Average_time
1	127.244	129.906	127.765	128.305	1	127.832	126.682	127.003	127.1723333
2	63.5174	64.2109	64.651	64.12643333	2	64.45	62.899	64.6309	63.9933
3	43.512	43.2895	42.889	43.23016667	3	46.416	46.104	46.3902	46.3034
4	31.5505	34.3147	33.005	32.95673333	4	32.9557	30.567	31.397	31.6399
5	26.014	25.4499	25.904	25.7893	5	25.5902	25.4212	25.0098	25.3404
6	20.65	21.9146	21.4301	21.33156667	6	21.8741	21.7809	20.6789	21.44463333
7	28.651	29.0832	29.224	28.98606667	7	25.3805	24.908	25.05	25.11283333
8	25.5825	24.344	25.003	24.9765	8	22.5597	22.8211	22.9201	22.76696667
9	23.771	22.981	23.4441	23.3987	9	21.1133	20.3366	21.251	20.9003
10	20.505	20.6068	19.987	20.36626667	10	20.449	19.359	20.3165	20.0415
11	18.898	18.475	15.988	17.787	11	18.9486	19.1	18.781	18.9432
12	15.004	17.2006	16.673	16.29253333	12	18.404	18.6138	18.72	18.57926667
					13	18.8714	17.45	18.8714	18.3976

SERIAL				
optimisa	T1	T2	T3	Average_time
0	128.12	127.5	127.003	127.541
1	71.2114	71.092	70.93	71.0778
2	58.2917	58.205	56.9981	57.8316
3	61.19	60.4929	61.9486	61.2105

CENTRE OF MASS

The centre of mass is at **(-0.09509, -0.16562, 49.64602)**

PERFORMANCE OF THE SERIAL, MPI AND OPENMP VERSIONS

The OPENMP and MPI versions have relatively the same performance. However they performed better than the SERIAL version.

For the log of time against the number of threads for the MPI and OPENMP, the graph starts high (Single process) and then progressively declines as the number of threads increases. However, when the number of threads is 7, there is an upward spike of the graph then followed by a smooth decline afterwards. This must be a result of either hyperthreading, load imbalance, synchronisation overhead and operating system scheduling. Of course hyperthreading has been used from thread 7 to thread 12.

For the average time against the number of threads, the plotted points is an exponential graph which indicates that as the number of threads increases, there is a corresponding decrease in the average time. This makes sense because as we introduce more threads, there is fewer

computation for each singular thread to do which thereby reduces the computational time. There is also an upward spike at thread 7 followed by a downward flattening of the curve afterwards.

For the Speedup against the number of threads, a straight line graph which shows that the speedup increases with increase in the number of threads. Then when the number of threads is 7, there is a sharp decline followed by a progressive increase once more. This must also be a result of the issues that were raised earlier (Gupta and Kumar, 1993).

For Efficiency, the graph starts high with an efficiency of 1 or 100%. We know that no process is 100% efficient. Therefore this must be as a result of approximations done during the calculations. Then there is a sudden little decline of efficiency at threads 3 and 4 of the OPENMP and MPI efficiency graphs respectively. This must be as a result of uneven load balancing. Then for hyperthreading, the efficiency dropped drastically and were below 0.7 or 70% for both MPI and OPENMP suggesting that as we add more threads, the system is less efficient in the distribution of work (Sharma and Kanungo, 2011).

CODE AMENDMENT

The following has been noticed to be the factors that negatively affects the computational time. However, due to want of time, It has not been effected in the source code.

1. The absense of pragma omp parallel for better loop parallelization.
2. The presence of nested “for loops” in the code which introduces several iterations thereby increasing computational time.

CONCLUSION

In conclusion, parallel computing significantly enhances the speed and efficiency of large-scale computations compared to traditional sequential processing. This report analysed three implementations of a Molecular Dynamics (MD) simulation program: a serial version, an OpenMP parallelized version, and an MPI parallelized version. The results showed that both OpenMP and MPI implementations outperform the serial version, although they present their own challenges, such as load balancing and synchronisation issues. The analysis revealed that increasing the number of threads improves performance up to a point, beyond which hyperthreading while reducing the computational time reduces efficiency as well. These findings underscore the importance of optimising thread management in parallel computing for maximum performance.

BIG DATA TASK

STATEMENT OF RESEARCH HYPOTHESIS

The research hypothesis which is the Alternate Hypothesis $H(1)$ states that Rides originating from Baylis Road, Waterloo station in 2014 were shorter than those from other stations.

STATEMENT OF THE NULL HYPOTHESIS

The corresponding Null hypothesis $H(0)$ states that Rides originating from Baylis Road, Waterloo station in 2014 were longer than those from other stations.

Following the pre-processing of the data, there are two groups which are the Average Duration of rides from Baylis road, Waterloo station and the Average duration of rides from Others. The Average duration of rides from Others is an aggregation of the other stations other than Baylis.

Month	Average Ride by Month Baylis	Average Ride by Month Others
01	960.2870813397129	1247.9157566495696
02	1064.6280991735537	1294.5294721724779
03	894.668721109399	1441.8672406565413
04	881.6007714561234	1531.8810665818007
05	1050.938673341677	1579.3038874589104
06	1110.615901455767	1569.8077658723566
07	1286.0921445144184	1570.6447931673715
08	1029.8165481093224	1653.3606386887209
09	1247.9394812680116	1414.1419520740462
10	879.4212765957446	1356.1646696133168
11	1048.9418777943367	1275.4024138845407
12	851.7014122394082	1412.0464056175222

In a nutshell, an independent sample t-test was carried on the data above. But since this test depends on other factors, other tests were carried out as well before the independent samples t-test.

TEST: A ONE TAILED INDEPENDENT SAMPLES T TEST

A one tailed independent sample t-test has been done to determine whether or not the research hypothesis is true. However, there are several conditions to be met before this test can be carried on the data.

Major conditions to be met before conducting and independent sample t test:

1. **Independency:** The samples from Baylis Road, Waterloo station and other stations must be independent of each other. This means that the duration of one ride should not influence the duration of another ride.
2. **Homogeneity of Variances:** The variances of the duration of rides in both groups should be approximately equal. This assumption is important because the t-test is sensitive to differences in variances between groups. The Levene's test has been used to test for this and the data was found to meet the criteria.
3. **Normality of data:** The durations of rides in both groups (Baylis Road, Waterloo station and other stations) should follow a normal distribution. This condition is especially important for smaller sample sizes like ours (typically $n < 30$). There is not an assumption of normal distribution. If the data does not follow the normal distribution, the test will not be effective (Ross and Willson, 2017). The Shapiro Wilk's test has been conducted to verify that the data follows this normal distribution.
4. **Data Continuity:** The data is continuous data. A different test other than the independent t samples test will be conducted if it were a categorical data.
5. **Random Sampling:** The data should be obtained through random sampling to ensure that the sample is representative of the population of interest. However in this case, all the rides from 2014 have been taken into consideration and this covers this criteria.
6. **The research hypothesis specifies a direction:** The critical area of the distribution is one-sided so that it is either greater than or less than a certain value but not both.

Summary of results for the Levene's test of homogeneity of variances:

Null hypothesis	Alternate hypothesis	Significance Level	Test Statistic	P-value
The variances of the groups (Baylis and Others) are equal	The variances of the groups (Baylis and Others) are not equal	A significance level of 5% or 0.05 was used for this test.	0.0084	0.9276

Conclusion: Since the p-value (0.9276) is greater than the significance level (0.05), we fail to reject the null hypothesis. Therefore, there is not enough evidence to suggest that the variances of the two groups (Baylis and Others) are different.

Summary of Results for the Shapiro Wilk's test for normality of data:

Null Hypothesis	Alternate hypothesis	Significance Level	Test statistic	P-value
The data (Baylis and Others) follows a normal distribution	The data (Baylis and Others) does not follow a normal distribution	A significance level of 5% or 0.05 was used for this test.	Baylis = 0.9164 Others = 0.9377	Baylis = 0.2576 Others = 0.4683

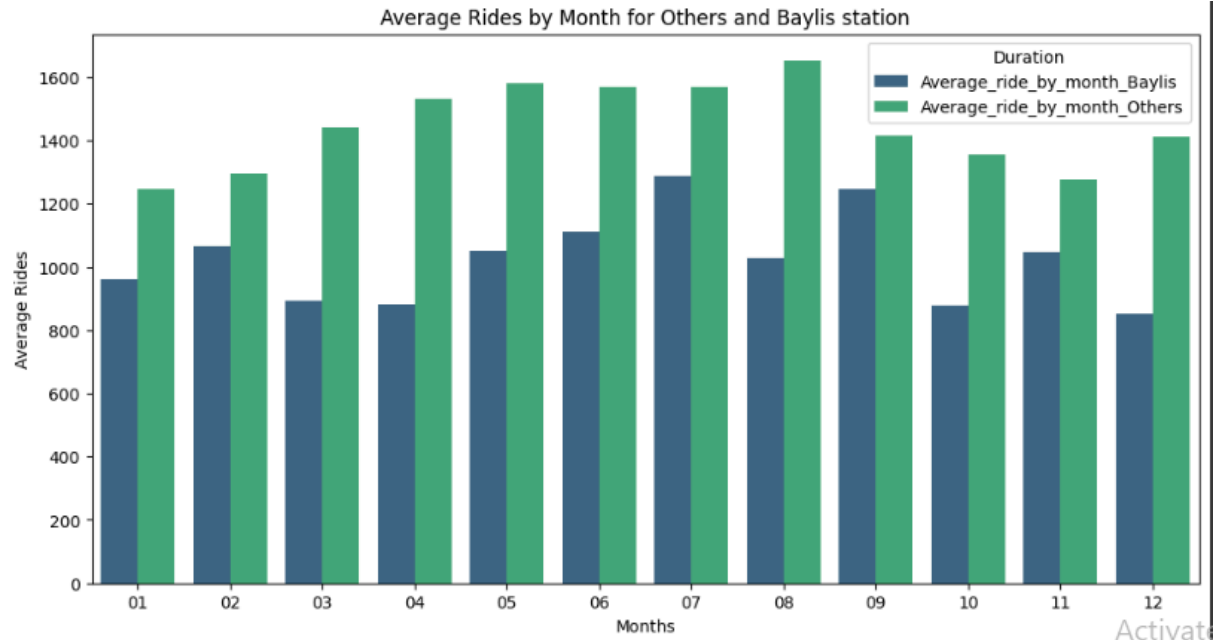
Conclusion: Since the p-values for Baylis and Others is greater than the significance level (0.05), we fail to reject the null hypothesis. Therefore, there is not enough evidence to conclude that the data significantly deviates from a normal distribution. Since Shapiro Wilk's test is sensitive to sample sizes (Shapiro et al., 1968), histograms were plotted as well to verify the normality of the groups.

INDEPENDENT SAMPLES T-TEST

Since the data meets all of these criteria, an independent t samples t-test was carried out.

Summary of results from the independent samples t-test:

Null Hypothesis	Alternate Hypothesis	Significance Level	T statistic	P-Value
Rides originating from Baylis Road, Waterloo station in 2014 were longer than those from other stations.	Rides originating from Baylis Road, Waterloo station in 2014 were shorter than those from other stations.	A significance level of 5% or 0.05 was used for this test.	-7.4327	9.7865×10^{-8}



Conclusion: Since the p-value is less than the significance level, there is a significant difference between the average duration of rides originating from Baylis Road, Waterloo station in 2014 compared to other stations. Therefore, we reject the null hypothesis.

This means that rides originating from Baylis Road, Waterloo station in 2014 were in actual sense significantly shorter than those from other stations at a significance level of 0.05.

THE ROLE OF BIG DATA IN ACHIEVING A MORE SUSTAINABLE GLOBAL ENVIRONMENT

Across numerous disciplines, especially in environmental sustainability, Big data analytics has shown to be an effective tool. Although, to achieve the aim of zero carbon emissions, this critical reflection will shine more lights on the social, legal, and ethical aspects of big data's role in creating a more sustainable global environment.

Social Aspect

Big data is very necessary for advocating environmental engagement and awareness on a social level. Parties involved may link communities, motivate eco-friendly behaviour, and gain deeper insights in the areas of environmental trends by using data from diverse sources, including social media, IoT devices, and satellite images. For instance, big data algorithms are used by platforms such as Ecosia to monitor deforestation and encourage users to participate in reforestation initiatives (Meet the team, 2024). However, considering marginalised communities may experience challenges to participate, it is crucial to resolve the digital divide in order to guarantee equal access to data-driven solutions (Liu et al., 2020).

Legal aspect

Some of the legal issues that arises when big data is being used are privacy issues , ownership, and regulatory compliance, especially when it comes to its usage in the area of environmental sustainability .To protect individual rights and regulate the gathering, storing , use of environmental data ,government have to make sure that very strong legal frameworks are in place . For example, the General Data Protection Regulation (GDPR) of the European Union sets rigorous guidelines for data processing and guarantees responsibility and transparency (European Commission, 2024). It is necessary to sync international legal norms to facilitate cross-border data exchange for global environmental initiatives.

Ethical aspect

The most noble effects of making use of big data for sustainable development are so noteworthy especially when it comes to topics of privacy, algorithm bias and consent. For example, as data collection rises also significantly, it is very crucial to protect highly sensitive information and only allow guaranteed informed permission.

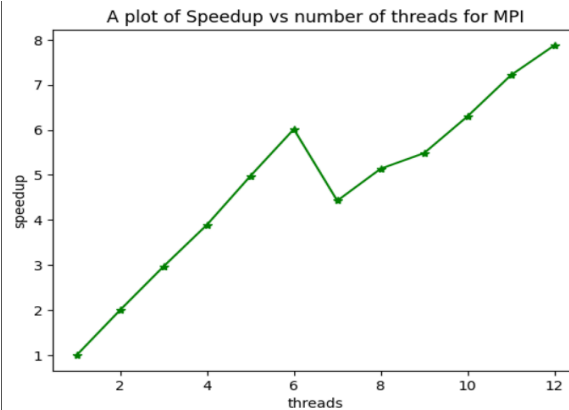
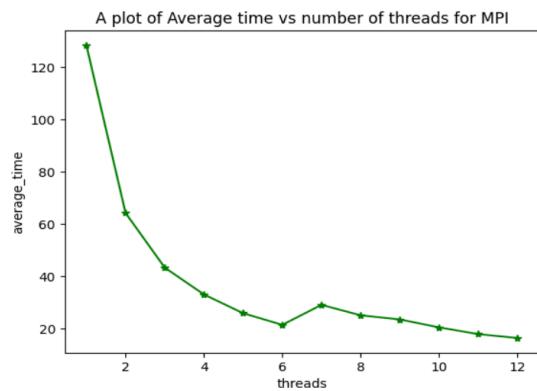
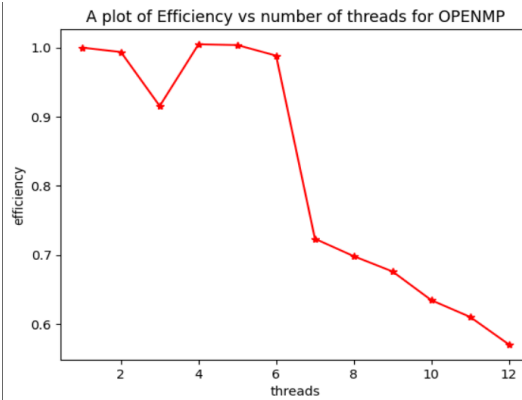
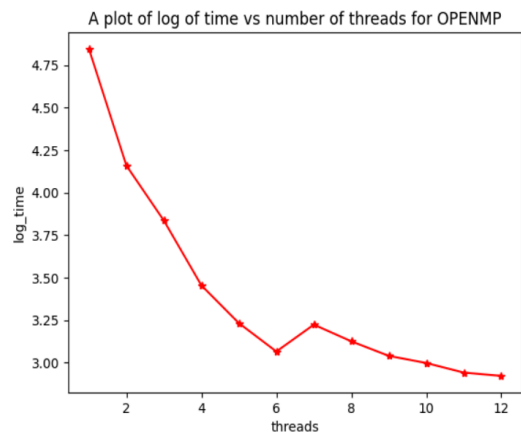
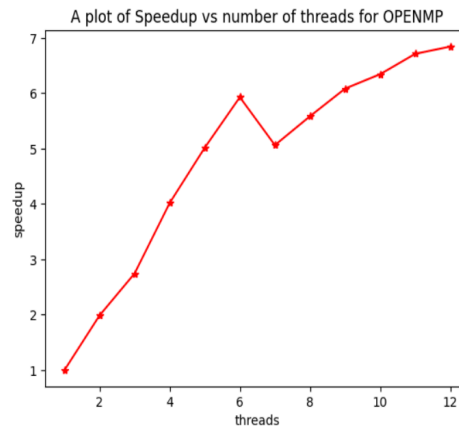
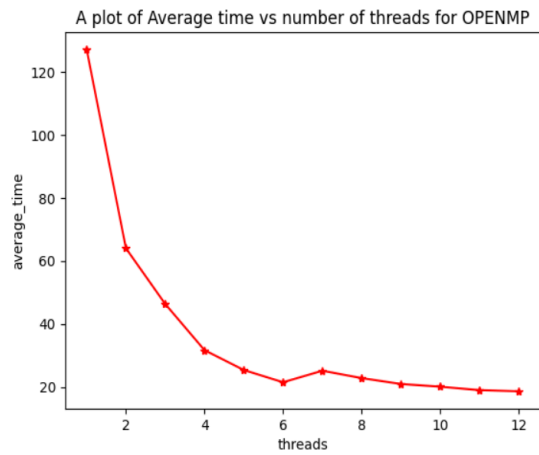
Furthermore, it is very important that the algorithms being used in decision making procedures be so explicit and responsible in order to steer clear from biased results and ethical concerns (Floridi et al., 2018).

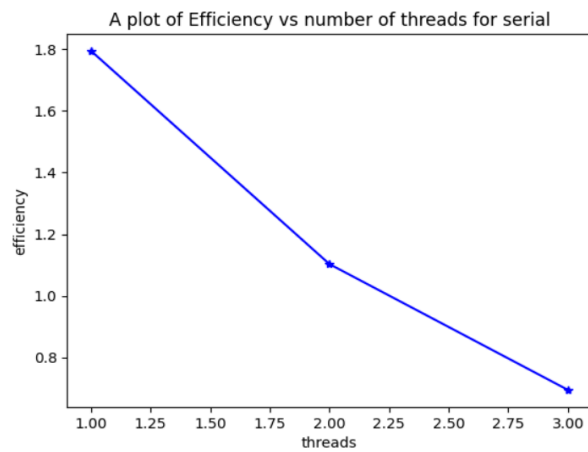
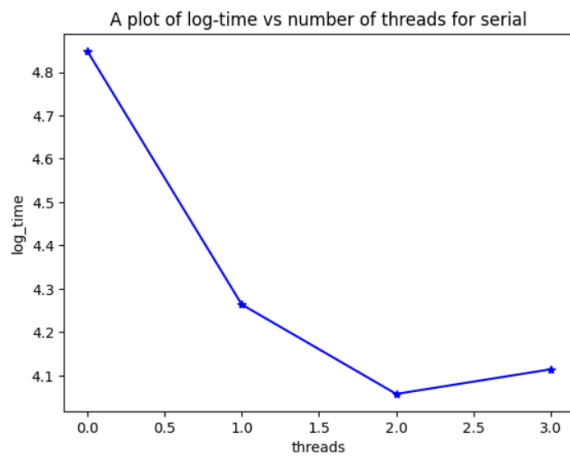
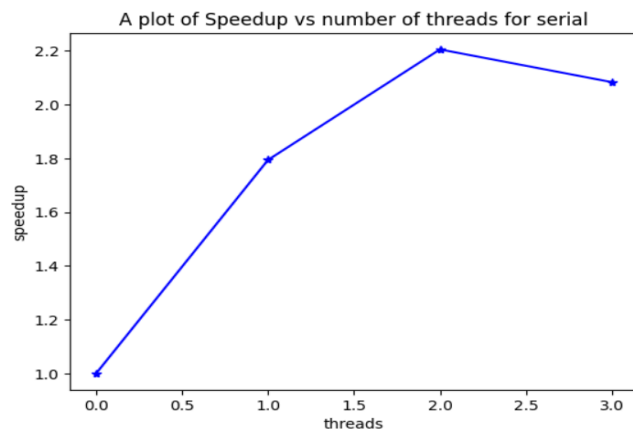
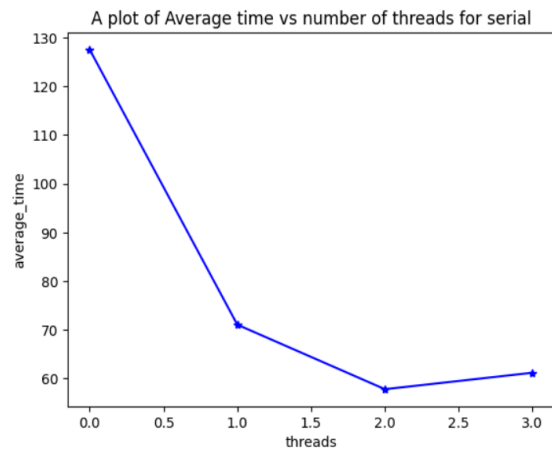
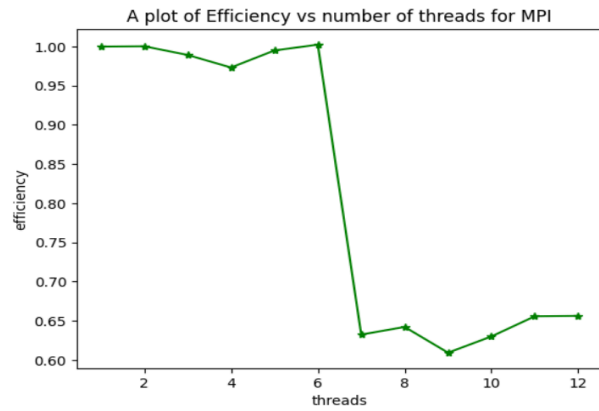
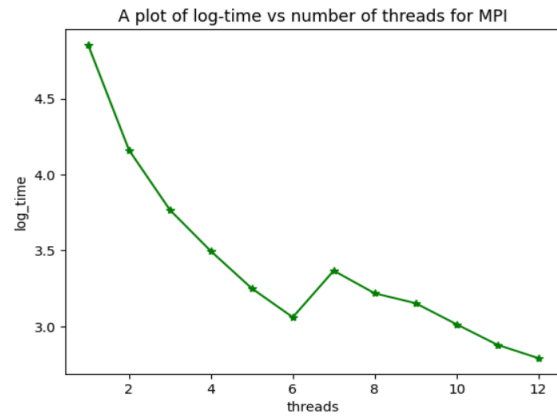
Development and implication of AI can be guided by ethical principles found in initiatives like IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE Ethics In Action in Autonomous and Intelligent Systems | IEEE SA, n.d.). However, to really deal with new

ethical complex problems, in the rapidly evolving field of big data analytics, persistent observation and alertness are essential.

In summary, big data serves multiple purposes that help foster a more sustainable environment, including social, legal, and ethical ones. Big data seems to present previously unprecedented opportunities to combat climate change and advance zero carbon emissions, although it also shows challenges with algorithms, fairness, data governance, and privacy protection. By establishing robust legal frameworks, ethical principles, and adopting inclusive social practices, stakeholders can effectively navigate the complex landscape of big data and utilise its transformative potential to promote environmental sustainability on an international level.

APPENDIX GRAPHS





REFERENCES

1. Czarnul, P., Proficz, J. and Drypczewski, K. (2020). 'Survey of methodologies, approaches, and challenges in parallel programming using high-performance computing systems.' Scientific Programming, 2020, pp.1-19.
2. Eager, D.L., Zahorjan, J. and Lazowska, E.D. (1989) 'Speedup versus efficiency in parallel systems.' IEEE transactions on computers, 38(3), pp.408-423.
3. European Commission. (2019) Data protection in the EU. European Commission. [Online] [Accessed on 22nd May 2024]
https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en.
4. Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689-707.
5. Gupta, A. and Kumar, V. (1993). 'Performance properties of large scale parallel systems.' Journal of Parallel and Distributed Computing, 19(3), pp.234-244.
6. IEEE Ethics In Action in Autonomous and Intelligent Systems | IEEE SA (n.d.). (2024) Ethics In Action | Ethically Aligned Design. [Online] [Accessed on 18th May 2024]
<https://ethicsinaction.ieee.org/>.
7. Liu, L., Song, X., and Zhang, Y. (2020). Bridging the Digital Divide: Towards Inclusive Big Data for Environmental Sustainability. Journal of Cleaner Production, 253, 119942.
8. Meet the team (2024) Ecosia.org. [Online] <https://info.ecosia.org/about>.
9. Ross, A. and Willson, V. L. (2017) Basic and Advanced Statistical Tests : Writing Results Sections and Creating Tables and Figures. Rotterdam: Sensepublishers.
10. Shapiro, S. S., Wilk, M. B. and Chen, H. J. (1968) 'A Comparative Study of Various Tests for Normality.' Journal of the American Statistical Association, 63(324) pp. 1343–1372.
11. Sharma, R. and Kanungo, P. (2011). 'Performance evaluation of MPI and hybrid MPI+ OpenMP programming paradigms on multi-core processors cluster.' In 2011 International Conference on Recent Trends in Information Systems (pp. 137-140). IEEE.