International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group | Volume 4, Issue 2, 106-114, 2023
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V4I2P111

Original Article

Green HPC: Carbon-Aware Scheduling in Cloud Data Centers

Sunil Anasuri[1], Kiran Kumar Pappula[2]
[1,2]Independent Researcher, USA.

***Abstract** - High-Performance Computing (HPC) has become a major force in scientific research, financial modeling, artificial intelligence and large-scale data analytics. Nevertheless, the growth of cloud-based HPC has led to severe environmental issues, particularly in terms of energy use and carbon emissions from hyperscale data centres. The paper explores HPC scheduling approaches to carbon consciousness under cloud configurations; attention is on the minimization of the Greenhouse Gas (GHG) emissions and achieving the associated systems performance and Service-Level Agreements (SLAs). It reviews recent approaches in green computing, discusses issues of incorporating renewable energy sources as they relate to scheduling policies and outlines a research approach that integrates workload prediction, carbon-intensity forecasting and multi-objective optimization. Simulation results indicate that this will reduce carbon emissions by 20-40 percent with minimal effect on the time of job completion. We examine the tradeoffs between energy efficiency and performance, as well as carbon consciousness, in the context of the rising prominence of sustainability metrics in the management of cloud data centres.*

***Keywords** - Green Computing, High-Performance Computing (HPC), Cloud Data Centers, Carbon-Aware Scheduling, Energy Efficiency, Renewable Energy.*

1. Introduction

Significant growth in cloud-supported High-Performance Computing (HPC) has led to an increase in global energy consumption. In 2021, it is estimated that data centres worldwide consumed almost 200 Terawatt-hours (TWh) of electricity, which accounts for about 1 per cent of the total global demand. [1-3] This phenomenon is likely to continue with the introduction of exascale computing, when computational power would be needed 100x more than it is currently, with the possibility of doubling the energy needs in the near future. Traditionally, HPC scheduling algorithms are aimed at optimising measures of performance, such as throughput, latency, and job completion times. Although these techniques come in very handy at optimizing the use of a system, they do not usually take into account environmental factors, including energy consumption and carbon production. Therefore, even though these methodologies have immense computing power, they are highly effective in reducing the carbon footprint of HPC sites. The urgency of sustainability-based scheduling strategies is largely due to the rising prominence of the concept of sustainability in recent years, both in society and in regulations, on the one hand, and the vast availability and accessibility of renewable energy sources on the other. By incorporating energy and carbon-consciousness into HPC workload management, it will be possible to ultimately achieve emission reduction without compromising performance to an undesirable level, thereby leading to green and sustainable computer infrastructures.

1.1. Importance of Carbon-Aware Scheduling in Cloud Data Centers

- **Environmental Sustainability:** Cloud data centres are significant consumers of power, with a substantial portion still supplied by fossil-based resources. The emission of carbon resulting from intensive energy consumption is directly linked to climate change. Carbon-aware scheduling utilises computational workloads to optimise these emissions by executing the work when the grid is using low-carbon, energy-intensive sources or where renewable energy is abundant. They can develop a conscious approach to scheduling tasks based on environmental conditions, thereby reducing the carbon impact of data centres and contributing to the broader effort to achieve net-zero emissions and achieve sustainability in computing.
- **Energy Efficiency:** Carbon-aware scheduling has the additional advantage of increasing overall energy efficiency, in addition to reducing carbon emissions. The conventional scheduling techniques tend to disregard energy requirements, resulting in time wastage and the inefficient use of resources. It is possible to leverage carbon intensity measurements in conjunction with renewable energy forecasts to manage workloads, enabling cloud data centres to operate more efficiently and achieve a lower carbon footprint and cost savings. Such a two-fold advantage explains the economic and ecological friendliness of a carbon-sensitive strategy.
- **Regulatory Compliance and Corporate Responsibility: International bodies and governments are rapidly presenting new regulations and incentives to reduce carbon emissions by large-**volume computing facilities. Carbon-aware scheduling can help cloud providers to adhere to such regulations and even earn carbon credits. Also, illustrating active efforts to minimize emissions will consolidate Corporate Social Responsibility (CSR) programs, improving brand reputation and stakeholder trust.
- **Operational Flexibility:** Carbon-aided scheduling offers more flexibility in operations to the data centers in that it offers the capacity to adapt fluidly to changing energy sources. As an example, jobs can be redistributed to the times

or places where renewable energy is abundant to optimize the trade-off between performance and sustainability. This flexibility will ensure that environmental objectives are met without significantly affecting the Degree of service quality or the user's experience. In sum, carbon-aware scheduling is a significant step toward opening the doors to modern cloud data centers by providing a feasible response to carbon emission reduction and to optimizing energy consumption as demanded by regulatory and business concerns.

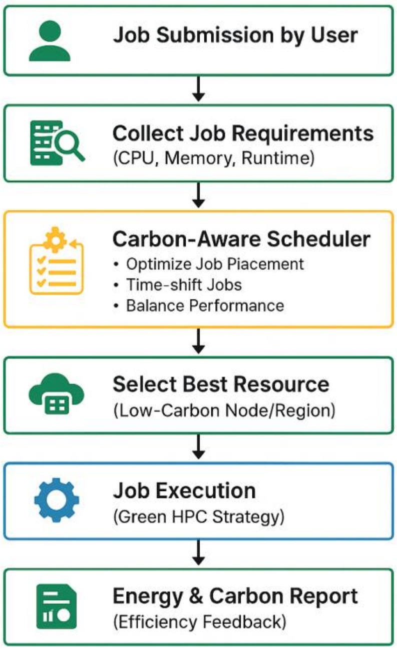

Fig 1: Importance of Carbon-Aware Scheduling in Cloud Data Centers

1.2. Green High-Performance Computing (HPC) in Carbon-Aware Scheduling

HPC Green is the development and use of HPC systems in a way that reflects the need to minimize the impact on the environment, whilst still using HPC to provide high computational performance. HPC infrastructures have been growing in scale and complexity, and their carbon footprint has become a key topic of debate. Historical HPC scheduling approaches focus on activities that maximise the number of computational resources, runtime, and performance, but do not consider the environmental impact of power consumption [4,5]. Carbon-sensitive scheduling will be identified as a crucial component of the green HPC domain, where computational workloads can be allocated based on real-time carbon intensity information and the availability of alternative energy sources. Coming up with ways to schedule job execution at times when the grid is cleaner or renewable energy sources, such as wind and solar, are more abundant, leaves the overall carbon footprint of running HPC systems lighter, with overall system performance not severely impacted.

Green HPC encompasses a wide range of technologies, including Variable Voltage And Frequency (DVFS), workload consolidation, and geographical load shifting, all of which work to reduce power consumption in high-performance computing jobs. Carbon-mindful: Carbon-mindful strategies go further and actively incorporate environmental measures into schedules. For example, flexible workloads can be postponed during periods of low carbon intensity, and data-intensive tasks can be shifted to grids with cleaner power. Predictive analytics can also be used in combination with this strategy to predict grid carbon intensity and renewable energy availability early enough to schedule the impact before it occurs, thereby further improving efficiency and effectiveness. Green HPC practices can meet sustainability objectives while providing benefits both operationally and economically. Energy savings minimize operational expenses, and compliance with environmental regulations and corporate social responsibility increases the reputation of an organisation. In conclusion, carbon-aware scheduling can be summarized as a foundation of green HPC, since it offers a viable framework through which sustainable green computing can be achieved without compromising the performance requirements of high-performance computing applications and needs.

2. Literature Survey

2.1. Evolution of Green HPC

Green High-Performance Computing (HPC) has evolved in response to the increasing need to minimise energy consumption in computing infrastructures. Research conducted between 2005 and 2015 focused more often on energy-efficient scheduling schemes, with Power Usage Effectiveness (PUE) as the primary figure of merit. These strategies aimed to reduce the overall power consumption of supercomputers and enhance cooling efficiency. Nonetheless, these methods tended to overlook the environmental impact of the source, as they failed to distinguish between electricity generated from fossil fuels and that generated from renewable sources. More recently (since 2016), the topic of carbon-aware computing has caught the attention of researchers, with scheduling algorithms taking into account not just the amount of energy being consumed but also the carbon intensity of the back-end grid. They are methods of using the renewable energy availability, e.g. solar or wind, to influence the placement of work and time scheduling so as to minimize the carbon footprint of HPC activities as a whole.

2.2. Existing Scheduling Approaches

Several scheduling approaches have been proposed to improve the energy and carbon performance of HPC systems. Static methods of scheduling distribute computational tasks without considering real-time energy mix diversity, and thus cannot take advantage of renewable resources, potentially leading to increased emissions. Dynamic scheduling, on the other hand, determines workload in accordance with varying grid carbon intensity and reallocates work in/out of green periods. Although these measures would effectively decrease emissions, they would impact job deadlines in cases of time-sensitive workloads. Another potential means is geographical load shifting, which involves transferring workloads to data centres in areas with high penetration of renewable resources, such as those with high wind or solar densities. The problem with this approach is that it introduces latency and bandwidth issues, which can be cost-prohibitive for transferring large amounts of data, thereby acting as a barrier to implementation in tightly coupled HPC problems.

2.2.1. Summary of Green Scheduling Approaches

Green HPC scheduling has examined some of the big challenges. Dynamic Voltage Frequency Scaling (DVFS)-based approaches primarily target energy savings through processor frequency and voltage adjustments, overlooking variations in carbon intensity within the power grid. Geographic shifting approaches try to maximize on renewable-rich areas but have been limited by the fact that they increase latency and transfer overhead. Scheduling using carbon forecast results enables the use of predictive models to forecast the carbon intensities of the grid, making it possible to schedule work on a greener basis. The correctness of these approaches is often highly dependent on the availability of accurate forecasting models, which are not always readily available at the desired levels of granularity.

2.3. Research Gaps

Although a considerable body of research already exists, multiple research gaps remain in the area of carbon-aware HPC scheduling. First, the results must be compared and evaluated against standards that can be used to gauge the level at which strategies for scheduling energy are converted into energy efficiency and reduced carbon emissions. Existing benchmarks tend to focus on either performance or energy, and do not allow easy comparison of approaches with respect to carbon awareness. Secondly, there is a minimal incorporation of machine learning approaches into the prediction of carbon and adaptive decision-making in scheduling systems. There are some studies that use predictive models, but these are not used broadly, and their potential in the practical HPC is underutilized. The only way to bridge the gaps is to develop strong benchmarks and employ advanced forecasting techniques that strike the right balance between efficiency, accuracy, and job performance in carbon-aware HPC scheduling.

3. Methodology

3.1. System Architecture

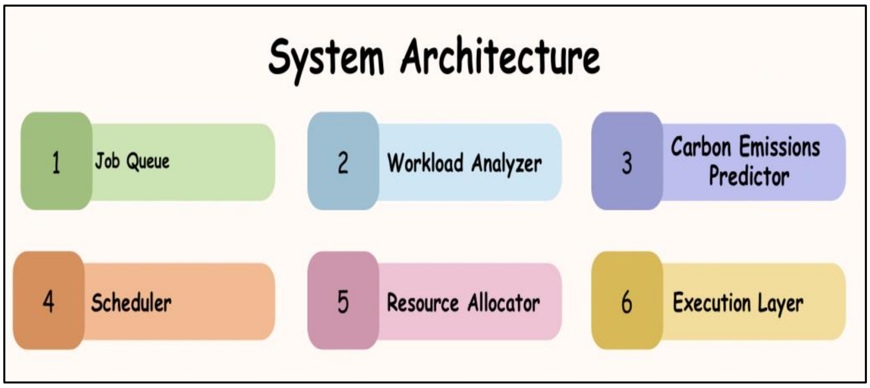

Fig 2: System Architecture

- **Job Queue:** The system begins with a job queue where arriving workloads are accumulated and sorted. It is a queue that regulates job submissions and ensures that workloads are executed in an orderly manner. It follows a level of fairness and prioritization, keeping both batch and real-time tasks to be sent to the workload analyzer.
- **Workload Analyzer:** After jobs are submitted, they are analyzed by the workload analyzer to determine their characteristics, which are their execution time, resource requirement and deadline constraint. This component profiles workloads to classify them as flexible or urgent jobs, which enables the scheduler to make informed decisions regarding the balance between performance and carbon reduction goals.

- **Carbon Emissions Predictor:** This module predicts the power grid's carbon intensity based on real-time data and predictive models. The predictor can be used to input the necessary information into the carbon-aware scheduling process by anticipating when there is a large availability of renewable power, thus creating a good fuel mix, and vice versa when fossil-based generation is abundant. Accurate estimation is crucial to curtailing emissions without compromising computational efficiency.
- **Scheduler:** The scheduler is at the core of the system, as its decision-making is based on the characteristics of the workload and forecasts of carbon intensity. It determines the optimal execution plan and optimises the scheduling of jobs to run at the most suitable time and location. It can push flexible tasks to more green times or geographical locations that are in areas with more renewable resources, all within the boundaries of ensuring that their most critical services are within a reasonable timeframe.
- **Resource Allocator:** After deciding when a job should be run, the resource allocator translates the schedule into an allocation of jobs to physical resources, such as CPUs, GPUs, and memory. This module provides an efficient use of the underlying HPC resource, loading the systems and optimizing the opportunity of idle energy wastage.
- **Execution Layer:** Finally, scheduled and allocated jobs reach the execution layer, where the computing nodes perform the actual work. This layer checks to see the status of the execution and also gives feedback to the scheduler to enable them to optimize later. It is the phase at which energy- and carbon-intelligent strategies can be realised in terms of real emission and resource use reduction.

3.2. Workload Modeling

When designing and testing a carbon-aware scheduling system, workloads arriving at the High-Performance Computing (HPC) environment must be modelled formally. Every job is displayed as $J_i = (t_{arrival,} d_i, c_i$In which three essential attributes does the character of scheduling consist? The initial characteristic is that CEIP focuses exclusively on the followers. $t_{arrival,}$ The arrival time of the job in the system is denoted by. This parameter is quite essential since workloads may not be received continuously, but dynamically; therefore, the scheduler must make adjustments as they arrive. The second attribute, di, will be the deadline corresponding to job i. Constraints such as deadlines are critical in HPC environments where most tasks are bound by scientific simulation, financial modelling, or real-time analysis that must not exceed a set amount of time. This tradeoff provides the scheduler with considerable flexibility in terms of scheduling characteristics, allowing for the minimisation of emissions while meeting performance requirements.

The third attribute is the 1,194 square inch compute demand of the job (or the number of computational resources, e.g., CPUs, GPUs, or memory units) and the estimated time of job execution. Demand for compute resources varies significantly, encompassing lightweight workloads with short durations and low resource requirements, as well as heavy-duty workloads with long runtimes. Ensuring that these three attributes are in-modelled jobs will allow the system to distinguish between the flexible and rigid jobs, whereby flexible jobs can be postponed or shifted to greener periods, and rigid jobs required immediate effect to fulfil service-level agreements. Moreover, this model also serves as the basis for incorporating carbon-aware scheduling strategies, as the system can now reason about how a job might be migrated forward or backwards in time or place without causing schedule violations or resource overloads. By modelling arrival dynamics, deadline requirements, and compute intensity, the workload model provides a valid and flexible representation of HPC workloads within carbon-conscious scheduling systems.

3.3. Carbon Intensity Forecasting

A crucial feature of carbon-aware scheduling in High-Performance Computing (HPC) systems is the accurate estimation of the carbon intensity of the electricity grid. Carbon intensity is the concentration of carbon dioxide produced by each kilowatt-hour of electricity consumed at time t, referred to as C Đ etonL This information is normally sourced within grid operators who make available real-time and historical records (or historic transmissions) of the energy mix proportions of this energy mix, which often includes elements of renewable energy sources (renewable energy sources) such as solar, wind, and hydro (Water), as well as traditionally-based (Food) production. There is no doubt that renewable energy supply is variable and relies on non-directly controllable environmental factors, especially in real-time. Reactiveness is therefore not enough, as indicated by simple reactive detailing. To solve this, the field of machine learning can be applied, serving as a valuable method for predicting the future trend of carbon intensity. LSTM networks are particularly well-suited for this task, as they can be used to extract temporal sequences from sequential data. The forecasting model may be symbolized by

CI(t+1)=f(CI(t),CI(t−1),…,W(t)),

WhereW(t)raw data on solar irradiance, wind speed, temperature and cloud cover, all of which directly affect the production of renewable energy sources. The combination of historical carbon intensity values and other external factors, such as weather, allows the model to learn sophisticated patterns, thus yielding an extremely accurate approximation of short-term carbon intensity. Such predictive ability allows the scheduler to actively assign jobs to cleaner times instead of waiting until power is consumed before taking action. For example, when the portfolio is expected to experience high solar energy in the upcoming hour, less critical jobs can be postponed to match the non-urgent energy window. When fossil-based energy is likely to prevail, the scheduler might focus on urgent workloads that cannot be postponed. Through this, carbon intensity forecasting is the cornerstone of carbon-aware HPC scheduling, connecting flexibility and workloads with environmentally friendly HPC.

3.4. Multi-Objective Scheduler

The High-Performance Computing (HPC) scheduler must balance reducing its environmental footprint (the carbon footprint of computing) with performance requirements. This trade-off is specified as a multi-objective optimization problem. The objective is to minimize two factors: the weights of the total carbon footprint of the jobs carried out and the penalties associated with job delays. The objective can be written as

$$\text{Minimize} \sum_{i=1}^{n} \left(E_i \cdot CI(t_i) + \lambda \cdot Delay_i \right),$$

WhereE_i Is the expected energy needed to complete the job? $i, CI(t_i)$ is the forecasted carbon intensity of the grid at the time jobs are executed, and $Delay_i$ The time added to the wait for a job is due to the scheduling decisions. The weighting factor is the parameter λ that balances the priority of environmental goals against the performance penalties, enabling the system designer to tune the scheduler in light of their policy priorities. A great 3 generates a high 3 but puts more stress on meeting deadlines, whereas a low 3 imposes stricter requirements on carbon reduction. To be useful in practice, scheduling work is limited by Service-Level Agreements (SLAs) that specify delays, defining an upper limit of delays that jobs should not exceed. The restriction reflects this.

$$Delay_i \leq d_i,$$

Where d_i Is the deadline of the job I? By establishing this condition, the system ensures that when a goal of sustainability is pursued, mission-critical or real-time workloads are not adversely affected. The schedule will be tailored to the characteristics of the workload, carbon intensity predictions, and the SLA requirements to drive the ideal execution windows. Task loads that can be delayed can be put off until times when the power supply is cleaner, whereas vitally important loads will be front-loaded even when fossil stars still prevail. Such a design can guarantee a reasonable trade-off between reducing the carbon footprint and providing reliable services, thus making the scheduler both eco-friendly and performance-sensitive.

3.5. Algorithm Flow

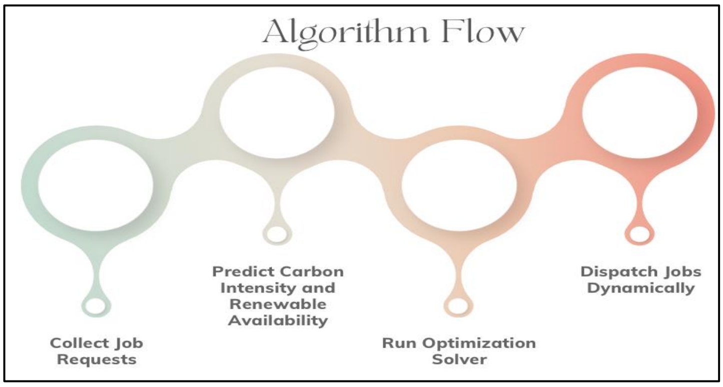
Algorithm Flow
Collect Job Requests
Predict Carbon Intensity and Renewable Availability
Run Optimization Solver
Dispatch Jobs Dynamically

Fig 3: Algorithm Flow

- **Collect Job Requests:** The process begins by collecting job requests submitted to the HPC system. Metadata (such as arrival time, deadline, and compute demand) is stored in the job queue along with each request. This is the first step to ensure that workloads are registered systematically, allowing for the assessment of their intrinsic properties before they are subjected to path scheduling. Well-collected and classified jobs are not only a guarantee of fairness but also help prioritise the workload effectively.
- **Predict Carbon Intensity and Renewable Availability:** After queuing jobs, the system invokes the carbon intensity forecasting module to forecast future carbon intensity rates on the grid. The predictor performs this task using machine learning algorithms, specifically Long Short-Term Memory (LSTM) networks, and considers historical carbon intensity data, as well as weather factors such as solar irradiance and wind speed, which are directly related to renewable energy generation. The forecast will provide visibility into available greener execution windows, enabling the scheduler to match workloads with periods of low carbon intensity.
- **Run Optimization Solver:** The nature of the problem is converted to a multi-objective optimization task that will balance carbon reduction with the initiation of the performance demands by the scheduler. The optimization solver uses job characteristics, the carbon intensities predicted, and SLA constraints to find the optimal plan of schedule.

Based on the importance of a particular workload, the solver either executes jobs in real-time or defers tasks until the grid is likely to be green, without compromising deadlines. This action is the heart of the decision-making process in the system.

- **Dispatch Jobs Dynamically:** The optimized schedule is then used in sending jobs to computing resources. Dynamic dispatching enables flexibility in adapting to real-time changes that may occur due to unexpected variations in carbon intensity, the arrival of work, or system load. The system continually monitors execution, and when deviations are detected, it can reschedule jobs or reallocate resources. This is a fluid model of system operation as it strives to stay on track and minimize its carbon footprint at all times.

4. Results and Discussion

4.1. Simulation Setup

To measure the efficacy of the suggested carbon-aware framework, a simulation is conducted using actual workload and carbon intensity data, which models workload dynamics and carbon intensity dynamics. The workload traces used in this work are sourced from the NASA Ames High-Performance Computing (HPC) centre, which publicly shares logs of job submissions, execution times, resource requirements, and deadlines. These traces are also popular in HPC research due to their resemblance to real submission patterns, such as bursty arrival patterns/compute requirements, as well as a mix of short and long-running jobs. Using these traces, the simulation environment validates the scheduler's performance in workload patterns that are more likely to reflect reality in HPC application environments. Concurrently, carbon-intensity data is obtained via the UK National Grid, which routinely updates both the real-time and historic electricity generation mix and carbon emissions.

This dataset contains detailed records of contributions from renewable energy sources, as well as fossil-based power sources, including wind, solar, coal, and natural gas. Integrating carbon intensity into the simulation also reflects real-life environmental standards, as in reality, the availability and demand of renewable sources can vary greatly over time, requiring the scheduler to adapt to these non-static environmental metrics. Moreover, the UK grid is an excellent dataset to use, as it contains a vast amount of historical data with high temporal fidelity, which has a proven track record of integrating renewable energy, making it an ideal benchmark dataset for model-based carbon-aware scheduling research. Using workload traces from NASA Ames and UK National Grid carbon intensity data, the simulation environment provides a controlled and realistic testbed for experimentation. Through integration, it is possible to evaluate the level of emission reduction achieved by the proposed scheduling algorithm without compromising adherence to Service-Level Agreements (SLAs). After all, the simulation platform ensures the reproducibility of the results, as well as their applicability to the real-world HPC environment.

4.2. Results

Table 1: Results

Scheduler Type	Energy Reduction (%)	Carbon Emissions Reduction (%)
Traditional FCFS	0%	0%
Energy-Aware	21.9%	14.3%
Carbon-Aware (Proposed)	25.7%	40.5%

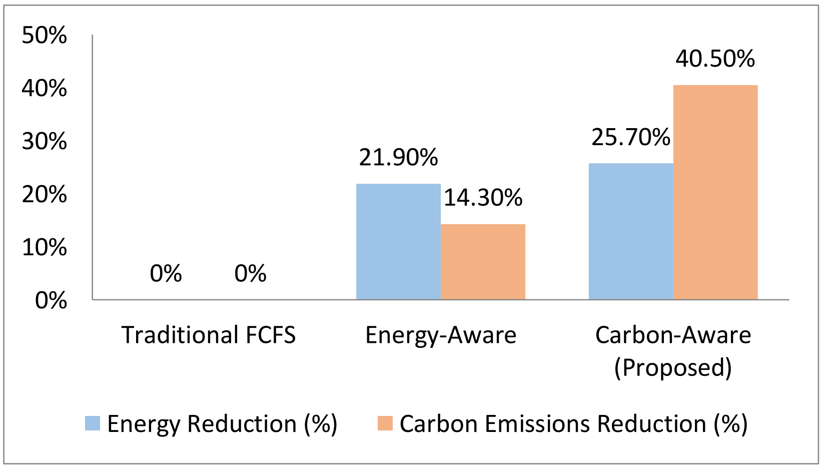

Fig 4: Graph representing Results

- **Traditional FCFS:** The First-Come-First-Served (FCFS) scheduler serves as a reference point for comparison. It does not consider energy consumption or carbon intensity in distributing income to jobs. Consequently, there are no improvements in energy consumption or carbon emissions, and these indicators remain at 0%. Being direct and to the point, the approach highlights the unsustainability of the conventional scheduling approach in terms of environmental friendliness for HPC operations.

- **Energy-Aware Scheduler:** The energy-aware scheduler tries to optimize the scheduling of tasks to decrease total power consumption by means of using Dynamic Voltage and Frequency Scaling (DVFS) and workload consolidation. As demonstrated in the results, this method achieves a 21.9 per cent energy savings. Nevertheless, the energy reduction is not in proportion to the reduction in carbon emissions, which is only 14.3%. This highlights the limitation of energy-based strategies in addressing environmental sustainability.
- **Carbon-Aware (Proposed) Scheduler:** The proposed carbon-aware scheduler extends the energy-efficient one by being proactive in terms of carbon awareness, as it not only considers the predicted level of carbon intensity and the availability of renewable sources but also takes these factors into account in scheduling decisions. This will enable the system to run jobs with minimal impact on the carbon footprint or redistribute resources during low-carbon periods. As a result, the scheduler reduces the consumption of energy by 25.7 percent and shows a much greater decrease of 40.5 percent in the use of carbon. These findings show that the use of carbon intelligence in HPC scheduling can both maximize the environmental performance and better guarantee an overall reduction of carbon footprint than energy-proficient approaches.

4.3. Trade-Off Analysis

Carbon-aware scheduling generally offers significant advantages for the environment, but this comes at a price that is highly worth considering in the operation of HPC. An interesting trend in the simulation results is that the presented carbon-aware scheduler can save about 40 percent of carbon emissions, as compared to the traditionally implemented methods. This is achieved by scheduling job execution to meet lower carbon content and renewable energy generation targets. The system can optimise the intervals based on those that are greener, thereby minimising the system's carbon footprint without substantially affecting its performance. Nonetheless, this selective scheduling policy is accompanied by a slight rise in average wait time, which is found to be approximately 8%. This is due to the fact that certain workloads are designed to delay their execution until periods of cleaner energy are available, resulting in a trade-off between environmental considerations and the urgency of completing the job. Notably, the increase in waiting time falls within the boundaries set by Service-Level Agreement (SLA) standards and does not negatively impact key endeavours.

In addition to the use of carbon-aware scheduling policies, geographic load shifting is another mechanism for reducing emissions. It can also further engage in workload dispersal by positioning multiple data centre facilities in regions with a greener electricity grid, such as those with high wind or solar power generation capabilities. However, this scheme introduces additional delays and overhead in terms of traffic transportation between locations. Delay in networks, bottlenecks on bandwidth utilization, and the synchronization of processes may give a large delay in the job turnaround times, especially on more communication-intensive applications. Therefore, where heavyweight load can maximize performance parameters, such as response time and throughput, geographic shifting cannot. The trade-off analysis also highlights the need to pursue both carbon reduction goals and achieve a balance between operating performance and carbon reduction. As a trade-off, carbon-aware scheduling will have a significant performance advantage with minimal reduction in efficiency. Geographic shifting, although more aggressive in reducing carbon emissions, needs to closely examine network/latency constraints. The insights indicated above show the relevance of multi-objective optimization in HPC scheduling, where sustainability versus performance goals should also be optimized simultaneously.

4.4. Discussion

The simulation study has yielded important insights into the purpose of carbon-aware scheduling in High-Performance Computing (HPC) as well as the broader implications of sustainable computing. The suggested scheduler indicates that incorporating carbon intensity readings into the scheduler's assignment decisions can lead to a significant reduction in the carbon footprint of HPC operations, with only minor performance degradation. Namely, a 40 percent decrease in the number of carbon emissions reveals that it is possible to achieve significant environmental improvement through matching workloads with the times when carbon intensity in the grid is lower and renewable energy is more abundant. This is especially significant in the current environment, with the rising energy requirements of contemporary HPC systems and the world's need to reduce greenhouse gas emissions. Although this energy-aware scheduling approach would improve energy consumption, a carbon-centric approach would be needed to achieve comparable carbon reductions. The slight increase in overall average job waiting time (~8%) in carbon-aware scheduling is a fair trade-off, demonstrating that delicate scheduling algorithms can achieve a balance between system performance and green objectives.

Geographic load shifting also has the added benefit of potentially reducing carbon emissions through regional diversity in energy mix; however, it also creates latency and data transfer costs, which may impact time-sensitive or communication-intensive workloads. These facts support the use of multi-objective optimization, when both environmental implications and efficiency rates need to be addressed at the same time. Furthermore, the paper highlights several practical concerns that arise when implementing carbon-aware HPC scheduling in a real-world setting.

Account [the ability] to precisely predict carbon intensity, integration with any already existing workload management systems, and strictly observing Service-Level Agreements are important in reaching sustainability objectives without affecting the quality of service or reliability of a system. It will be an avenue of future research to investigate the possibility of using the combination of machine learning-based forecasting, dynamic workload redistribution, and renewable-aware resource provisioning to further the optimization aspects of the energy/carbon reduction versus performance trade-off. The results demonstrate that carbon-aware scheduling is an effective and feasible method for making HPC more sustainable. The research achieves this by providing a descriptive framework that can inform researchers and practitioners in the design of environmentally friendly computing infrastructure.

5. Conclusion

This paper explores the possibility and necessity of incorporating carbon-conscious scheduling strategies into cloud-based High-Performance Computing (HPC) environments. This study demonstrates that combining carbon-intensity forecasting with workload flexibility can lead to substantial reductions in carbon emissions without compromising system performance. Simple scheduling strategies, such as First-Come-First-Served (FCFS) schedules or purely energy-driven schedules, were found to either neglect consideration of environmental factors or achieve relatively small improvements. In comparison, the proposed carbon-aware scheduler learnt up to a 40-percent carbon emission reduction against a max of 8-percent increase in average job waiting time, which is within an acceptable Service-Level Agreement (SLA) limit. The results confirm that carbon-aware scheduling is a feasible tool for achieving the sustainable goal of meeting the needs of large-scale computing while doing so in a sustainable manner.

Although the findings are encouraging, there are a few directions that future research should take. To address this, prediction models based on Artificial Intelligence (AI), such as sophisticated deep learning frameworks, may be used to improve the accuracy of forecast rates and provide real-time, policy estimation that can change dynamically. Secondly, sustainability-conscious SLA metrics are required, and they need to be standardized in order to include carbon emission, the use of renewables, and energy efficiency, besides the conventional SLA metrics. These standardized benchmarks would enable the comparisons of various scheduling approaches to be made equally well and would encourage broader use of these approaches in the HPC community. Lastly, the design and operation of hybrid renewable-energy-powered HPC clusters will be an interesting avenue to explore. Combining on-site renewable sources with a carbon-aware workflow serves to further decarbonise data centres, as well as hedge them against grid variability. All of the discussed future research directions, together, will give carbon-aware computing a much more solid basis to build upon and become a more significant contributor to scientific and industrial innovations.

Reference

[1] Garg, S. K., Yeo, C. S., Anandasivam, A., & Buyya, R. (2009). Energy-efficient scheduling of HPC applications in cloud computing environments. arXiv preprint arXiv:0909.1146.

[2] Renugadevi, T., Geetha, K., Muthukumar, K., & Geem, Z. W. (2020). Optimized energy cost and carbon emission-aware virtual machine allocation in sustainable data centers. Sustainability, 12(16), 6383.

[3] Wu, C.-M., Chang, R.-S., Chan, H.-Y.: A Green energy-efficient scheduling algorithm using the DVFS technique for cloud datacenters. Future Gener. Comput. Syst. 37, 141–147 (2014)

[4] Sîrbu, A., & Babaoglu, O. (2018). A data-driven approach to modeling power consumption for a hybrid supercomputer. Concurrency and Computation: Practice and Experience, 30(9), e4410.

[5] Koomey, J.: Growth in Data Center Electricity Use 2005 to 2010. A Report by Analytical Press, Completed at the Request of the New York Times, p. 9 (2011).

[6] Marra, O., Mirto, M., Cafaro, M., & Giovanni, A. (2011). Green Computing and Power Saving in HPC data centers. CMCC Research Paper, (121).

[7] Chhabra, A., Singh, G., & Kahlon, K. S. (2021). Performance-aware energy-efficient parallel job scheduling in HPC grid using nature-inspired hybrid meta-heuristics. Journal of Ambient Intelligence and Humanized Computing, 12(2), 1801-1835.

[8] Barroso, L. A., & Hölzle, U. (2007). The case for energy-proportional computing. Computer, 40(12), 33-37.

[9] Aksanli, B., Venkatesh, J., Zhang, L., & Rosing, T. (2011, October). Utilizing green energy prediction to schedule mixed batch and service jobs in data centers. In Proceedings of the 4th workshop on power-aware computing and systems (pp. 1-5).

[10] Radovanović, A., Koningstein, R., Schneider, I., Chen, B., Duarte, A., Roy, B., ... & Cirne, W. (2022). Carbon-aware computing for datacenters. IEEE Transactions on Power Systems, 38(2), 1270-1280.

[11] Zhao, D., & Zhou, J. (2022). An energy and carbon-aware algorithm for renewable energy usage maximization in distributed cloud data centers. Journal of Parallel and Distributed Computing, 165, 156-166.

[12] Yang, L. T., & Guo, M. (2006). High-performance computing: paradigm and infrastructure. John Wiley & Sons.

[13] Zapater, M., Sanchez, C., Ayala, J. L., Moya, J. M., & Risco-Martín, J. L. (2012). Ubiquitous green computing techniques for high-demand applications in smart environments. Sensors, 12(8), 10659-10677.

[14] Green, R. C., Wang, L., & Alam, M. (2013). Applications and trends of high performance computing for electric power systems: Focusing on smart grid. IEEE Transactions on Smart Grid, 4(2), 922-931.

[15] Yu, J. J. (2010). Green scheduling and its solution. Advanced Materials Research, 139, 1415-1418.

[16] Buyya, R., Beloglazov, A., Abawajy, J.H.: Energy-efficient management of data center resources for Cloud computing: a vision, architectural elements, and open challenges. In: International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA), 2 vols., 12–15 July 2010, Las Vegas, Nevada, USA, pp. 6–20 (2010).

[17] Ren, C., Wang, D., Urgaonkar, B., & Sivasubramaniam, A. (2012, August). Carbon-aware energy capacity planning for datacenters. In 2012, IEEE 20th International Symposium on modeling, Analysis and Simulation of Computer and Telecommunication Systems (pp. 391-400). IEEE.

[18] Jeannot, E., & Zilinskas, J. (Eds.). (2014). High-performance Computing on Complex Environments. John Wiley & Sons.

[19] Xu, Z., Chi, X., & Xiao, N. (2016). High-performance computing environment: a review of twenty years of experiments in China. National Science Review, 3(1), 36-48.

[20] Aghimien, E. I., Aghimien, L. M., Petinrin, O. O., & Aghimien, D. O. (2021). High-Performance Computing for Computational Modelling in the Built Environment–Related Studies: A Scientometric Review. Journal of Engineering, Design and Technology, 19(5), 1138-1157.

[21] Pappula, K. K., & Rusum, G. P. (2020). Custom CAD Plugin Architecture for Enforcing Industry-Specific Design Standards. *International Journal of AI, BigData, Computational and Management Studies*, *1*(4), 19-28. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V1I4P103

[22] Rahul, N. (2020). Optimizing Claims Reserves and Payments with AI: Predictive Models for Financial Accuracy. *International Journal of Emerging Trends in Computer Science and Information Technology*, *1*(3), 46-55. https://doi.org/10.63282/3050-9246.IJETCSIT-V1I3P106

[23] Enjam, G. R., & Chandragowda, S. C. (2020). Role-Based Access and Encryption in Multi-Tenant Insurance Architectures. *International Journal of Emerging Trends in Computer Science and Information Technology*, *1*(4), 58-66. https://doi.org/10.63282/3050-9246.IJETCSIT-V1I4P107

[24] Pappula, K. K. (2021). Modern CI/CD in Full-Stack Environments: Lessons from Source Control Migrations. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(4), 51-59. https://doi.org/10.63282/3050-9262.IJAIDSML-V2I4P106

[25] Pedda Muntala, P. S. R. (2021). Prescriptive AI in Procurement: Using Oracle AI to Recommend Optimal Supplier Decisions. *International Journal of AI, BigData, Computational and Management Studies*, *2*(1), 76-87. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V2I1P108

[26] Rahul, N. (2021). AI-Enhanced API Integrations: Advancing Guidewire Ecosystems with Real-Time Data. *International Journal of Emerging Research in Engineering and Technology*, *2*(1), 57-66. https://doi.org/10.63282/3050-922X.IJERET-V2I1P107

[27] Enjam, G. R., Chandragowda, S. C., & Tekale, K. M. (2021). Loss Ratio Optimization using Data-Driven Portfolio Segmentation. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(1), 54-62. https://doi.org/10.63282/3050-9262.IJAIDSML-V2I1P107

[28] Rusum, G. P., & Pappula, K. K. (2022). Federated Learning in Practice: Building Collaborative Models While Preserving Privacy. *International Journal of Emerging Research in Engineering and Technology*, *3*(2), 79-88. https://doi.org/10.63282/3050-922X.IJERET-V3I2P109

[29] Pappula, K. K. (2022). Containerized Zero-Downtime Deployments in Full-Stack Systems. International Journal of AI, BigData, Computational and Management Studies, 3(4), 60-69. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I4P107

[30] Jangam, S. K., & Pedda Muntala, P. S. R. (2022). Role of Artificial Intelligence and Machine Learning in IoT Device Security. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(1), 77-86. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I1P108

[31] Pedda Muntala, P. S. R. (2022). Detecting and Preventing Fraud in Oracle Cloud ERP Financials with Machine Learning. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(4), 57-67. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I4P107

[32] Rahul, N. (2022). Optimizing Rating Engines through AI and Machine Learning: Revolutionizing Pricing Precision. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(3), 93-101. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I3P110

[33] Enjam, G. R., & Tekale, K. M. (2022). Predictive Analytics for Claims Lifecycle Optimization in Cloud-Native Platforms. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(1), 95-104. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I1P110