# Toward Optimal Carbon-Aware Scheduling of Server Replacement

Ankur Sharma, Rishabh Tamhane, Anshul Gandhi
Stony Brook University, NY, USA
{ankursharma, rtamhane, anshul}@cs.stonybrook.edu

## ABSTRACT

Server replacement scheduling decisions can significantly impact carbon emissions in data centers. However, commonly used periodic strategies are often sub-optimal in terms of total carbon emissions. In this article, we formulate the carbon-aware server replacements problem mathematically and investigate the structure of the optimal replacement policy. Based on our theoretical analysis, we then propose a simple adaptive online policy that yields significantly lower carbon emissions than periodic strategies and achieves near-optimal results across a variety of simulated scenarios.

## 1 Introduction

Data centers employ numerous servers, and these servers are often replaced periodically due to wear-and-tear, to avoid ensuing failures, and also to exploit newer, better technology [18, 12]. The newer technology is typically more efficient, even in terms of carbon efficiency, including both operational and embodied carbon. For example, if we consider the configuration of a high-end PowerEdge R760 server and consider different CPU microarchitecture generations (e.g., Coffee Lake, Cascade Lake, Rocket Lake released in 2017, 2019, and 2021, respectively), we get estimated embodied carbon costs of 996, 888.1, and 794.1 kgCO2e [3]. As such, replacing servers can impact the carbon footprint of a data center.

However, server replacement as a tool for carbon efficiency has not been well explored. Current practices are to replace servers periodically [10, 4] or when they fail [14, 15, 13]. There is thus an opportunity here to reduce carbon costs by scheduling replacements based, at least partially, on trends in operational and embodied carbon.

We develop an initial model to assess the impact of server replacement on total carbon cost of a server and then theoretically analyze the carbon optimization problem. We obtain exact optimal results in some cases, which we use to design an adaptive heuristic that can be easily applied for generic cases. Numerical evaluation results show that our adaptive policy outperforms periodic policies substantially, and, importantly, is usually with 1% of the offline optimal solution.

## 2 Server Replacement: Problem Setup & Analysis

Consider a server that is purchased at time $t_0 = 0$ and is replaced at times $t_1 < t_2 < ... < t_n$. Let the time horizon of analysis be $T$; so, $T > t_n$. For simplicity, we consider time to be in units of years. We consider that a server purchased at time $t$ incurs a (one-time) embodied carbon cost of $E(t)$ and its operational carbon cost is $O(t)$ per year that it is in operation. In the R760 example mentioned earlier, setting $t = 0$ for 2017, we get $E(0) = 996, E(2) = 888.1, E(4) = 794.1$. The $E()$ and $O()$ functions/curves can take any form (e.g., linear, exponential, etc.); we expect the costs to usually reduce over time due to, for example, technological advances.

The key challenge is that ***the future*** $E()$ ***and*** $O()$ ***values are unknown***, so the replacement decision at any time $t$ (whether to replace now or not) must be made using only the information ($E()$ and $O()$ values) available until time $t$. The objective then is ***to find, in an online manner, the replacement periods*** $t_1, t_2, .., t_n$ ***to minimize the total carbon cost***:

$$C = \sum_{i=0}^{n} \left\{ E(t_i) + O(t_i) \cdot (t_{i+1} - t_i) \right\}, \qquad (1)$$

where $t_0 = 0$ and $t_{n+1} = T$. The cost function, $C$, can be extended as needed; for example, by multiplying $O()$ with the carbon intensity, we can covert $C$ to the SCI (Software Carbon Intensity) metric [7, 9, 8].

We acknowledge the above formulation is a starting point for a carbon model on server replacements. While it captures the essential features of the problem and yields important insights into the structure of the carbon-optimal server replacement schedule, it does not account for several real-world complexities. These include, for example, the finite lifetime and failure rates of servers, the presence of multiple server options in the market with differing carbon profiles, and the possible degradation of operational efficiency over a server's lifetime (for instance, due to wear or changing carbon intensity of energy sources). Additionally, performance requirements and constraints are not explicitly modeled. Nevertheless, as we will show in the following, the formulation provides a valuable analytical foundation and yields heuristics that can inform the design of more practical replacement policies. There is also potential to extend the framework to incorporate some of these complexities in future work (see Section 4).

**A general result:** We first establish an interesting property that must hold in an optimal replacement schedule for any $E()$ and $O()$ curves. For this result, we assume $t$ is continuous as it allows us to use calculus, and the continuous results can be shown to typically carry over to the discrete setting. Consider any three successive replacement times in an optimal schedule: $t_{j-1}, t_j, t_{j+1}$. We fix $t_{j-1}$ and $t_{j+1}$ and consider the change in total cost as we vary $t_j$. Collecting terms in Eq. (1) that depend only on $t_j$, the total carbon cost, in terms of $t_j$, can be written as

*CarbonMetrics 2025* Stony Brook, NY

$C(t_j) = O(t_{j-1}).t_j + E(t_j) + O(t_j).(t_{j+1} - t_j) + \text{constant}$. Differentiating $C(t_j)$ and setting it to zero gives us, with some rearrangement of terms:

$$t_{j+1} - t_j = \frac{O(t_{j-1}) - O(t_j) + E'(t_j)}{-O'(t_j)} \qquad (2)$$

For specific $E()$ and $O()$ curves, the above result lets us derive interesting relations between successive replacement intervals in an optimal schedule. For example, if embodied cost and operational costs are both decreasing linearly, say $E(t) = E_0 - s \cdot t$ and $O(t) = O_0 - k \cdot t$, then Eq. (2) gives us $(t_{j+1} - t_j) = (t_j - t_{j-1}) - \frac{s}{k}$; so, successive replacement intervals in an optimal schedule must decrease by a constant amount in this case.

**The case of linear $O()$ and constant $E()$:** The above analysis when applied to the case of linear $O()$ and a constant $E() = E$, tells us that *optimal replacements must happen periodically.* Assume that the optimal replacement interval is $x$. For ease of analysis, we set $T = (n+1) \cdot x$ (this can be made precise by a careful discrete analysis). Then, from Eq. (1), we get the total cost as a function of $x$ as $C(x) = \sum_{i=0}^{n}[E + (O_0 - k \cdot i \cdot x) \cdot x] = \frac{E \cdot T}{x} + O_0 \cdot T - \frac{1}{2} \cdot k \cdot (T - x) \cdot T$. To minimize the cost, we set $C'(x) = 0$, giving us:

$$\text{replacement period (x)} = \sqrt{2 \cdot E/k} \qquad (3)$$

We note that similar results have been derived in unrelated settings in the Operations Research community [17, 11]. However, we have not come across results similar to Eq. (2) or the above proof of optimality of the periodic solution. Similarly, Bashroush et al. have provided a single-decision-point analysis, based on energy consumption estimates, showing that appropriate server replacement can reduce energy consumption for a realistic workload [1, 2]. However, the analysis in that work does not consider online replacements over time, lacks a formal optimization framework, and focuses on energy rather than carbon.

**The case of arbitrary $O()$ and $E()$:** An exact analysis for the case where $E()$ is not a constant but is also decreasing linearly can be similarly carried out, though it does not result in a closed-form expression. For more general functional forms of $E()$ and $O()$, even the expressions for the total cost can get complicated.

As such, *finding the optimal replacement periods in an online manner for arbitrary carbon cost curves remains a challenge.*

**Our proposed adaptive replacement policy:** For an arbitrary $O()$ curve, at time $t$, its instantaneous slope can be obtained from the functional form as $O'(t)$ or by approximating it as $O'(t) \approx O(t) - O(t-1)$. Then, using Eq. (3) as the basis, we can determine the next replacement period as:

$$\text{adaptive replacement period} = \sqrt{2 \cdot E(t)/|O'(t)|} \qquad (4)$$

In effect, this policy locally approximates the $O()$ curve as linear and $E()$ curve as a constant. While the policy need not be optimal for general $O()$ and $E()$ curves, it provides a simple and practical heuristic to obtain replacement schedules in an online and adaptive way, taking into account changing $E()$ and $O()$ values. As we show next, our easy-to-use policy performs surprisingly well.

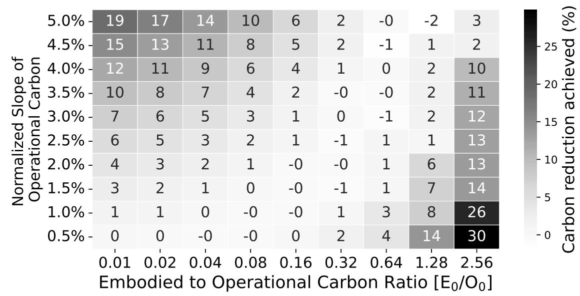


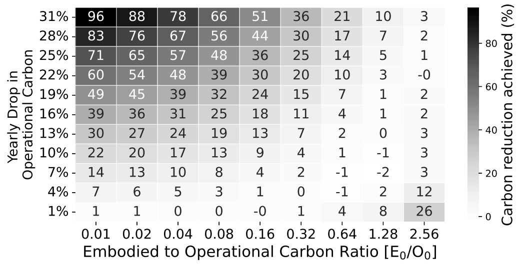

Figure 1: Percentage reduction in total carbon achieved by our adaptive policy over 5-year periodic replacement under linear (yearly) carbon cost curves.

# 3 Evaluation Results and Insights

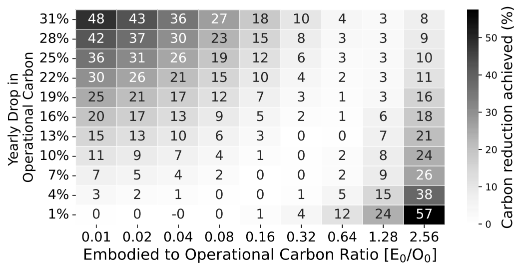
We now present numerical results to evaluate the carbon cost savings afforded by our adaptive policy. We analyze for a $T = 20$ year horizon and consider $E()$ and $O()$ curves that are linear or exponential. We compare the total carbon cost over 20 years incurred by our policy with that incurred by (1) a *periodic* policy that replaces the server every 5 years (unless stated otherwise), based on replacement periods reported in prior work [10, 4]; and (2) the impractical *offline optimal* policy, obtained via dynamic programming, wherein the entire $E()$ and $O()$ curves are known in advance.

**Linear $E()$ and $O()$ curves:** Figure 1 presents the carbon cost reduction afforded by our adaptive policy over the 5-year periodic replacement policy as a heatmap; darker shades represent higher reductions, with the values in each cell indicating the percentage reduction. On the x-axis, we vary the ratio of initial embodied to operational costs ($E_0/O_0$) from 0.01 (meaning embodied cost at $t = 0$ is 1% of the annual operation cost at $t = 0$) to 2.56, on a log scale. The y-axis represents the slope of the linearly decreasing $O()$ curve, denoted as a percentage of the initial operational carbon cost, $O_0$. That is, y-axis represents $k/O_0$ as a percentage, referring to the $O(t) = O_0 - k \cdot t$ linear model from Section 2. The slope of the linearly decreasing $E$ curve (denoted as $s$ in Section 2) is set to half the slope of the $O()$ curve; that is, $s = 0.5 \cdot k$. We also experimented with $s/k$ ranging from 0.1–0.9 but the results were qualitatively similar. The ranges for x- and y-axis values have been selected based on realistic $E()$ and $O()$ estimates provided by vendors for their various server and laptop models [16, 3]. We set $O_0 = 3,000$ kgCO2e based on available numbers for servers [5, 6]; we find that the cost reduction results for a given x- and y-axis value pair are invariant to $O_0$ (as they cancel out when comparing policies).

The top-left region of Figure 1 corresponds to situations where the embodied costs are low and the annual reduction in operational costs $(O(t) - O(t-1))$ is large. In such situations, replacements incur little embodied cost but provide substantial carbon savings; as such, waiting 5 years between replacements is sub-optimal, and we see up to ***19%*** improvement using our method. The bottom-right region corresponds to the situation where the embodied costs are high and the change in annual operational costs is small. Replacements should thus be infrequent, and the periodic policy incurs high costs (up to ***30%*** higher




(a) Compared to 5-year periodic replacements.




(b) Compared to 3-year periodic replacements.

Figure 2: Percentage reduction in total carbon achieved by our adaptive policy over (a) 5-year and (b) 3-year periodic replacement under exponential carbon cost curves.

compared to our adaptive policy) performing multiple, unneeded replacements.

Over the entire range of results shown in Figure 1, our adaptive policy provides an average carbon cost reduction of about 4.5% over the 5-year periodic policy. Importantly, the *carbon cost difference between our adaptive policy and the offline optimal policy is **within 0.6%***, averaged over all cases considered in Figure 1.

**Exponential $E()$ and $O()$ curves:** We now consider exponentially decreasing carbon cost curves as $E(t) = E_0 \cdot e^{-s \cdot t}$ and $O(t) = O_0 \cdot e^{-k \cdot t}$. The results, illustrated in Figure 2(a), show similar trends as Figure 1, but the improvements afforded by our policy are more pronounced. Our policy outperforms the 5-year periodic policy by up to ***96%*** in the top-left region (low embodied but high operational drop), and by up to ***26%*** in the bottom-right region (high embodied and low operational drop). Further, the carbon cost of our policy is ***within 0.35%*** of that achieved by the offline optimal when averaged over all cases shown in Figure 2(a). These promising results are not limited to comparing with a 5-year periodic policy. Figure 2(b) shows our improvements compared to a 3-year periodic policy. The 3-year periodic policy, by replacing more frequently than the 5-year policy, does slightly better in the top-left region but performs even worse in the bottom-right region. Over the entire range of results shown in the figures, the average improvement our policy provides over the 5- and 3-year policies is 12% and 21.6%, respectively.

We see that our adaptive policy provides greater improvements in the exponential case. This is because, under the exponential cost curves, it is better to replace more frequently initially (when the costs are dropping rapidly) and replace infrequently later on (when costs have plateaued). While our adaptive policy indeed follows this pattern, the static periodic policy is unable to do so.

# 4 Conclusions and Future Work

We proposed a mathematical model to analyze the total carbon cost of server replacements and explored structural properties of the carbon-optimal server replacement policy. A general result was established, and through the analysis of a special case, we derived closed-form expressions that informed the design of a simple, adaptive heuristic. This heuristic was shown to perform well in example settings—achieving results close to the offline optimal and, in certain scenarios, substantially outperforming periodic replacement policies.

Looking ahead, future work can extend the analysis along several dimensions. A deeper investigation of the model may yield additional structural insights or more refined heuristics. In particular, it would be valuable to examine more complex carbon cost trajectories—including those with abrupt shifts. Incorporating more realistic system-level considerations, such as finite server lifetimes, performance constraints, degradation in efficiency over time, and the availability of multiple server options at each decision point, would enhance the practical relevance of the model and the policies it informs. Moreover, realistic replacement decisions often involve objectives beyond carbon, such as dollar costs. These typically also exhibit a one-time fixed component (such as purchase cost) and a running cost (such as maintenance costs), and can be naturally incorporated into our formulation by appropriately combining them with the embodied and operational cost curves.

# Acknowledgment

This work was supported by NSF grants CCF-2324859, CNS-2214980, CNS-2106434, and CNS-1750109.

# 5 References

[1] R. Bashroush. A comprehensive reasoning framework for hardware refresh in data centers. *IEEE Transactions on Sustainable Computing*, 3(4):209–220, 2018.

[2] R. Bashroush, N. Rteil, R. Kenny, and A. Wynne. Optimizing server refresh cycles: The case for circular economy with an aging moore's law. *IEEE Transactions on Sustainable Computing*, 7(1):189–200, 2022.

[3] Boavizta. Boavizta — environmental-footprint-data. `https://github.com/Boavizta/environmental-footprint-data`, 2025.

[4] R. D. Caballar. Data Center Hardware Refresh Cutback by Microsoft — What's Next?, 2022.

[5] Dell Technologies. Product Carbon Footprint POWEREDGE R250. `https://www.delltechnologies.com/asset/en-us/solutions/infrastructure-solutions/industry-market/pcf-report-poweredge-r250-15g.pdf`, 2022.

[6] Dell Technologies. Product Carbon Footprint POWEREDGE R350. `https://www.delltechnologies.com/asset/en-us/solutions/infrastructure-solutions/industry-market/`

pcf-report-poweredge-r350-15g.pdf, 2022.

[7] A. Gandhi, K. Ghose, K. Gopalan, S. Hussain, D. Lee, D. Liu, Z. Liu, P. McDaniel, S. Mu, and E. Zadok. Metrics for sustainability in data centers. In *Proceedings of the 1st Workshop on Sustainable Computer Systems Design and Implementation (HotCarbon'22)*, San Diego, CA, USA, July 2022. USENIX.

[8] G.-D. Hoffmann and V. Majuntke. Improving carbon emissions of federated large language model inference through classification of task-specificity. In *Proceedings of the 3rd Workshop on Sustainable Computer Systems Design and Implementation (HotCarbon'24)*, Santa Clara, CA, USA, 2024.

[9] S. Köhler, L. Wenzel, A. Polze, B. Herzog, H. Hofmeier, M. Vögele, and T. Hönig. Carbon-aware memory placement. In *Proceedings of the 2nd Workshop on Sustainable Computer Systems Design and Implementation (HotCarbon'23)*, Boston, MA, USA, 2023.

[10] J. Lyu, M. You, C. Irvene, M. Jung, T. Narmore, J. Shapiro, L. Marshall, S. Samal, I. Manousakis, L. Hsu, P. Subbarayalu, A. Raniwala, B. Warrier, R. Bianchini, B. Schroeder, and D. S. Berger. Hyrax: Fail-in-Place server operation in cloud platforms. In *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, pages 287–304, Boston, MA, USA, 2023.

[11] S. Mueller. *Investment and Production*. Harvard University Press, Cambridge, MA, 1961.

[12] S. Mueller, M. E. Soper, and B. Sosinsky. *Upgrading and repairing servers*. Pearson Education, 2006.

[13] I. Narayanan, D. Wang, M. Jeon, B. Sharma, L. Caulfield, A. Sivasubramaniam, B. Cutler, J. Liu, B. Khessib, and K. Vaid. SSD Failures in Datacenters: What? When? And Why? In *Proceedings of the 9th ACM International Conference on Systems and Storage*, SYSTOR '16, Haifa, Israel, 2016.

[14] B. Schroeder, R. Lagisetty, and A. Merchant. Flash reliability in production: The expected and the unexpected. In *Proceedings of the 14th USENIX Conference on File and Storage Technologies (FAST 16)*, pages 67–80, Santa Clara, CA, USA, 2016.

[15] B. Schroeder, E. Pinheiro, and W.-D. Weber. Dram errors in the wild: A large-scale field study. In *Proceedings of the Eleventh International Joint Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '09, page 193–204, New York, NY, USA, 2009. Association for Computing Machinery.

[16] M. Stutz, S. O'Connell, and J. Pflueger. Carbon footprint of a dell rack server. *2012 Electronics Goes Green 2012+*, pages 1–5, 2012.

[17] G. Terborgh. *Dynamic Equipment Policy*. McGraw-Hill, New York, 1949.

[18] G. Wang, L. Zhang, and W. Xu. What can we learn from four years of data center hardware failures? In *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 25–36. IEEE, 2017.