Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

SIMULATING JOB REPLICATION VERSUS ITS ENERGY USAGE

Vladimir Marbukh
Brian Cloteaux

National Institute of Standards and Technology
Information Technology Laboratory
100 Bureau Drive
Gaithersburg, MD 20799, USA

ABSTRACT

Due to the proliferation of computers in all aspects of our lives, the energy and ecological impacts of computing are becoming increasing important. Some of the transformative algorithms of recent years generate huge amounts of carbon dioxide, potentially damaging the environment. We have developed a set of simulations for understanding the trade-offs between distributed computing and its carbon impact. We briefly describe our current work and our future research aiming at finding practical algorithmic solutions.

1 INTRODUCTION

Our modern society has been, in large part, built upon computing. Energy generation and its uses have become interwoven into our society. Unfortunately, in recent years we have seen the ecological consequences of carbon-based energy sources. Thus, it has become increasingly important to look for ways to mitigate this carbon-based dependency while still supporting the technological advances in our society.

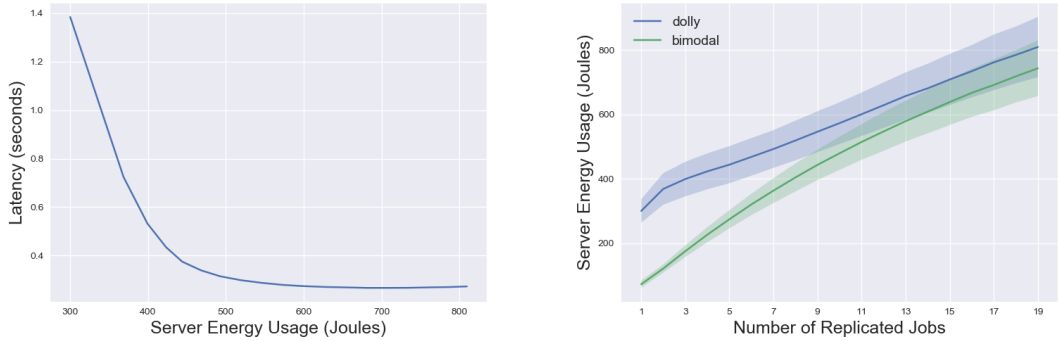
One problem in sustainable computing is to lessen the use of carbon-based energy in highly distributed computations (Agarwal et al. 2021). This is a challenging problem, because the intermittency of green versus non-green energy sources due to external conditions such wind velocities and the amount of sunlight. Mitigating this problem can be approached by scheduling based on the amount of green energy available at different server centers and storing green energy for later use. Due to the complexity of the problem, simulations may play an important role in finding practical solutions.

2 SIMULATING THE ENERGY USAGE DUE TO JOB REPLICATION

Resource usage versus performance trade-offs occur naturally in distributed algorithms. Distributing sub-calculations based on resource constraints (such as time or attached peripherals) is a well-studied field. Our initial approach has been to introduce the additional constraint of carbon expenditure, and then study the trade-off between the time (both computation time and latency) and the energy usage.

We are now focusing on the trade-offs and energy limitations in the application of computational replication in mitigating intermittency (Gowrisankaran et al. 2016). Replication covers a set of strategies where jobs can be submitted to multiple servers at the same time for processing. By using replication, jobs can potentially benefit in two ways: first, protection against server crashing and slowdowns and, second, providing a quicker response time since the response time of the job is essentially dictated by the shortest server queue in which a replicated job is placed. Unfortunately, by allowing multiple copies of a job to start, we are obviously expending energy resources. We are interested in maintaining the advantages of job replication while mitigating the additional energy usage.

To study this trade off, we have developed a simulation framework allowing us to examine the effects of various models and policies on the usage of sustainable energy for distributing jobs across different server farms. These models are build by combining features such as redundancy policies and server slowdown distributions (Gardner et al. 2017) with communication and server energy usage models (Boru et al. 2015). Our simulation allows us to evaluate different scenarios for distributing replicated jobs under various energy policies. Example simulations results produced by this software package for various scenarios are shown in Figure 1.



(a) Trade-off between average latency and average server energy usage.

(b) Average energy usage versus number of replicated jobs using differing server slowdown models.

Figure 1: Plot 1a shows the tradeoff of energy use versus latency time for a cancel-on-completion redundancy policy (Gardner et al. 2017). Plot 1b shows how differing server slowdown models affect server energy usage. The error bars show one standard deviation for 10000 samples.

3 CURRENT AND FUTURE WORK

Currently, we are developing cost-based algorithms for balancing energy usage with performance (Palomar and Chiang 2006). These mechanisms will allow for individual jobs to "purchase" job replication based on their resource needs and the current costs associated with individual servers. These costs could fluctuate based on the availability of energy supplies for the server. The aim is to develop non-centralized algorithmic policies for balancing energy usage with latency (Marbukh and Mills 2008).

REFERENCES

Agarwal, A., J. Sun, S. Noghabi, S. Iyengar, A. Badam, R. Chandra, S. Seshan, and S. Kalyanaraman. 2021, November. "Redesigning Data Centers for Renewable Energy". In *Proceedings of the Twentieth ACM Workshop on Hot Topics in Networks*, HotNets '21, 45–52. New York, NY, USA: Association for Computing Machinery.

Boru, D., D. Kliazovich, F. Granelli, P. Bouvry, and A. Y. Zomaya. 2015, March. "Energy-efficient data replication in cloud computing datacenters". *Cluster Computing* 18(1):385–402.

Gardner, K., M. Harchol-Balter, A. Scheller-Wolf, M. Velednitsky, and S. Zbarsky. 2017, August. "Redundancy-d: The Power of d Choices for Redundancy". *Operations Research* 65(4):1078–1094. Publisher: INFORMS.

Gowrisankaran, G., S. S. Reynolds, and M. Samano. 2016, August. "Intermittency and the Value of Renewable Energy". *Journal of Political Economy* 124(4):1187–1234. Publisher: The University of Chicago Press.

Marbukh, V., and K. Mills. 2008, April. "Demand Pricing & Resource Allocation in Market-Based Compute Grids: A Model and Initial Results". In *Seventh International Conference on Networking (icn 2008)*, 752–757.

Palomar, D., and M. Chiang. 2006, August. "A tutorial on decomposition methods for network utility maximization". *IEEE Journal on Selected Areas in Communications* 24(8):1439–1451. Conference Name: IEEE Journal on Selected Areas in Communications.