Energy-Aware Scheduling for Serverless Computing

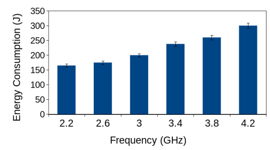
Michail Tsenos, Aristotelis Peri, Vana Kalogeraki
Athens Univ. of Economics and Business
Athens, Greece
{tsemike,ariperi,vana}@aueb.gr

1 Introduction

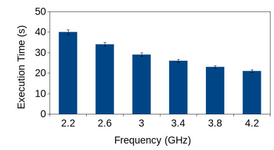
Serverless computing paradigm is a popular cloud computing model that has gained considerable traction in recent years. This scheme offers an event-driven framework for easily developing and scaling cloud-based applications, while also lowering execution costs. One aspect of such systems that has received limited research attention is the energy consumption of these systems during application execution. One way to deal with this issue is to schedule the function invocations in an energy-efficient way. However, efficient scheduling of applications in a multi-tenant environment, like FaaS systems, poses significant challenges. The trade-off between the server's energy usage and the hosted functions' performance requirements need to be taken into consideration. In this work, we propose an Energy Efficient Scheduler for orchestrating the execution of serverless functions so that it minimizes energy consumption while it satisfies the applications' performance demands. Our approach considers real-time performance measurements and historical data and applies a novel DVFS technique to minimize energy consumption. Our detailed experimental evaluation using realistic workloads on our local cluster illustrates the working and benefits of our approach

2 Motivation

Recent studies have shown that the global data-centre electricity consumption in 2022 was estimated at 1-1.3% of global final electricity demand. Also, the cost of powering servers housed in datacenters comprises about 30% of the total cost of ownership. Although modern CPUs are more energy efficient than the previous generations, further scheduling decisions can result in even better energy savings without violating the performance of the running workloads. Modern CPUs support DVFS which is a power management technique used to dynamically adjust the operating voltage and frequency based on workload demands. By intelligently modulating these parameters, DVFS can optimize energy efficiency without sacrificing performance. This technique allows CPUs to run at lower frequencies and voltages based on workload requirements saving power directly and indirectly. Directly it decreases the energy consumption of the CPU and indirectly, by reducing the heat decimated by the CPUs, auxiliary systems in the datacenters such as cooling pumps and air conditioning units can be turned off. In Fig.1a we can observe the co-relation between the power consumption and the performance of a SHA256 benchmark application running with different CPU frequencies on the same machine.



(a) Energy consumption on various CPU frequencies



(b) Execution time on various CPU frequencies

3 Methodology

The main research question that we try to answer is how to lower the energy consumption in clusters that host serverless functions without violating their deadlines. To do that we developed two components. The first component is Processor Management and Scheduling Utility which is a component that runs on each cluster node and is responsible for (i) adjusting in real-time the CPU clock frequency and (ii) reporting in real-time to the energy-efficient scheduler crucial information that it needs for each node such as the current CPU clock frequency, temperature and power values. The second component is the energy-efficient scheduler that determines the number of instances and the placement of serverless function containers in a cluster of nodes. It determines the most suitable number of instances and their frequency configuration based on a queuing theoretic model and by utilizing historical data from previous runs of the same job. If no such data is available, a profile run is performed and the required number of instances is found by using a prediction model that utilises known throughput curves of functions with similar CPU utilisation. The proposed Energy Efficient Scheduler (EES) is a hybrid greedy scheduler that adapts to varying loads. It prioritizes nodes with matching CPU frequencies under low loads and during high loads, the scheduler evaluates its scheduling decisions' impact on energy cost, deploying jobs accordingly to optimize resource utilization in the available nodes. This approach balances performance and energy efficiency in serverless environments. Our experimental evaluation has shown that EES is from 14% up to 28% (Fig. 1b)superior to its competitors in terms of energy savings without violating any deadline.

Acknowledgments

This research has been financed by the Hellenic Foundation for Research and Innovation (HFRI) under the 3rd Call for HFRI PhD Fellowships (Fellowship Number: 6812), by the European Union through the EU ICT-48 2020 project TAILOR (No. 952215), the H2020 AutoFair project (No. 101070568) & the Horizon Europe CoDiet project (No. 101084642).