

Green Coding

Software energy optimization by understanding the impact of programming languages



Mohammed Chakib Belgaid

Supervisors: Pr. Romain Rouvoy
Pr. Lionel Seinturier

University of Lille

This dissertation is submitted for the degree of

Doctor of Philosophy

EDSPI

September 2022

I would like to dedicate this thesis to my loving parents ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Mohammed Chakib Belgaid
September 2022

Acknowledgements

And I would like to acknowledge ...

Abstract

This is where you write your abstract ...

Contents

List of Figures	xv
List of Tables	xix
1 Motivation	1
1.1 Motivation	2
1.2 Research Contributions	2
1.3 Publications	2
2 State of the Art	3
2.1 Benchmarking	5
2.1.1 Threats & Challenges	5
2.1.2 Proposed Solutions	6
2.2 Energy Measurement	7
2.2.1 Hardware Tools	8
2.2.2 Software Tools	10
2.2.3 Hybrid Tools	12
2.3 Programming Languages & Performances	12
2.3.1 Energy	12
2.3.2 Python	13
2.3.3 JVM	13
2.3.4 benchmarking	14
2.4 Energy Variation	14
2.4.1 Studying Hardware Factors	14
2.4.2 Mitigating Energy Variations.	16
2.5 Conclusion	16

I Benchmarking Programming Technologies	19
3 A Benchmarking Protocol to Measure Software Energy Consumption	21
3.1 Motivation	21
3.2 Reproducibility within the context of energy	22
3.2.1 Docker Vs. Virtual Machine	23
3.2.2 Docker & Energy	23
3.2.3 Conclusion	25
3.3 Improving Accuracy by Taming Energy Variations	26
3.3.1 Experimental Setup	26
3.3.2 Analysis	29
3.3.3 Docker & Accuracy	29
3.3.4 Experimental Guidelines	43
3.3.5 Threats to Validity	46
3.3.6 Conclusion	46
3.3.7 Extension	47
4 Energy Footprint of Programming Languages	49
4.1 Investigating Remote Procedure Call Frameworks	50
4.1.1 Introduction	50
4.1.2 Research Questions	50
4.1.3 Experimental Protocol	50
4.1.4 Results & Findings	52
4.1.5 Threats to Validity	56
4.1.6 Conclusion	56
4.2 Investigating Web Application Frameworks	56
4.2.1 Introduction	57
4.2.2 Comparaison of Web Frameworks	59
4.2.3 Energy Efficiency in Software Engineering	59
4.3 Experimental Protocol	61
4.3.1 Measurement Context	61
4.3.2 Workload	64
4.3.3 Key Performance Metrics	67
4.3.4 Extension	68
4.4 Results & Findings	68
4.4.1 Overall all statistics	69
4.5 Threats to Validity	71

II Optimizing Application Runtimes	77
5 The Impact of Python Runtime on Energy Consumption	79
5.1 Introduction	79
5.2 Motivation	80
5.2.1 Python Popularity	80
5.2.2 Python Gluttony	81
5.2.3 Python Limits	82
5.3 Green Web Development	83
5.3.1 life cycle of a request in Django	83
5.3.2 Python Insights	87
5.3.3 Python & Multiprocessing	90
5.3.4 Python & Machine Learning	91
5.4 Results & Findings	92
5.5 Python Interpreters	94
5.5.1 Runtime Classification	97
5.6 experimental protocole	98
5.6.1 measurement context	98
5.6.2 Metrics	101
5.6.3 tests preparation	103
5.6.4 Extension	103
5.7 Results and finding	104
5.7.1 python and binarry operations	108
5.8 conclusion	109
6 The Impact of Java Virtual Machine on Energy Consumption	113
6.1 Introduction	113
6.1.1 Goal	113
6.2 Experimental Protocol	114
6.2.1 Measurement Contexts	114
6.2.2 Workload	114
6.2.3 Metrics and Measurement	115
6.2.4 Extension	117
6.3 Experiments & Results	118
6.3.1 Energy Impact of JVM Distributions	118
6.3.2 Energy Impact of JVM Settings	122
6.4 Threats to Validity	130

6.5 Tools and contributions	131
6.6 Conclusion	131
7 Discussion and Conclusion	133
7.1 Conclusion	133
7.2 Summary of Contributions	133
7.3 Limitations and Challenges	133
7.4 Future Work	133
Bibliography	135

List of Figures

2.1	electrical cost comparasion between two cpus	4
2.2	The proposed decomposition of an experiment [14]	8
2.3	CPU energy variation for the benchmark CG	15
2.4	the spiral method of energy optimization	17
3.1	Different Methods of Virtualization	23
3.2	energy consumption of Idle system with and without docker [?]	24
3.3	Execution time & energy consumption of Redis with and without Docker [?]	25
3.4	Topology of the nodes of the cluster Dahu	27
3.5	Comparing the variation of binary and Docker versions of aggregated LU, CG and EP benchmarks	30
3.6	Energy variation with the normal, sleep and reboot modes	31
3.7	STD analysis of the normal, sleep and reboot modes	32
3.8	Energy variation when disabling the C-states	33
3.9	Energy variation considering the three cores pinning strategies at 50 % workload	34
3.10	C-states effect on the energy variation, regarding the application processes count	36
3.11	OS consumption between idle and when running a single process job	38
3.12	The correlation between the RAPL and the job consumption and variation .	39
3.13	Energy consumption STD density of the 4 clusters	42
3.14	Energy variation comparison with/without applying our guidelines	44
3.15	Energy variation comparison with/without applying our guidelines for STRESS-NG	45
3.16	Benchmarking protocol	48
4.1	Experimental software architecture	51
4.2	Experimental software architecture	57
4.3	Experimental software architecture	58

4.4	Experimental software architecture	59
4.5	Experimental software architecture	60
4.6	Experimental software architecture	61
4.7	Experimental software architecture	63
4.8	Experimental software architecture	64
4.9	Experimental software architecture	65
4.10	Architecture	68
4.11	Spearman Rank Correlation between different metrics	70
4.12	correlation of latency and number of requests per second for a single query	71
4.13	Average power consumption for the Single query test	72
4.14	total request vs average power consumption for the Single query test (size of circles represents the number of clients)	72
4.15	total request vs average power consumption for the multiple queries test (Size of circles represents the size of the query)	73
4.16	total request vs average power consumption for the update queries test (Size of circles represents the size of the query)	73
4.17	total request vs average power consumption for Fortunes test (Size of circles represents the number of clients)	74
4.18	total request vs average power consumption for plainText test (size of circles represents the number of clients)	74
4.19	total request vs average power consumption for JSON Serialization test (size of circles represents the number of clients)	75
5.1	PYPL Popularity of Programming Languages [?].	79
5.2	Energy consumption of a recursive implementation of Tower of Hanoi pro- gram in different languages [95]	80
5.3	Use cases of Python (source: JetBrains).	82
5.4	Request-Response life cycle in Django	84
5.5	Tree representation of the energy consumption of a single request in django (naive version)	84
5.6	Energy behavior resulting from data access strategies.	86
5.7	Tree representation of the energy consumption of a single request in django (naive version)	87
5.8	Energy consumption of Python multiprocessing depending on the number of exploited threads.	88
5.9	Comparison of the energy consumption of different Python loops.	89
5.10	Comparison of the energy consumption of different methods to convert a list.	89

5.11	energy behaviour based on multiprocessing	90
5.12	accuracy based on epoch	92
5.13	cumulative energy consumption vs accuracy	93
5.14	evolution of average gpu power based on epoch	93
5.15	cumulative energy of fast10 benchmark within different configurations . . .	94
5.16	Python interpeters	100
5.17	powerapi architecture	102
5.18	102
5.19	Dendogram of the different implementations	106
5.20	Different interpererts optimisation score	110
5.21	energy consumption of different impelementations using Bit Operation benchmarks (Joule)	111
5.22	Summary of the binary operations for different python implementations . .	111
6.1	Target scope of DACAPO and RENAISSANCE benchmarks.	117
6.2	Energy consumption evolution of selected JVM distributions along versions.	118
6.3	Energy consumption of the HotSpot JVM along versions.	119
6.4	Energy consumption comparison across Java benchmarks for HOTSPOT, GRAALVM & J9.	120
6.5	Power consumption of Scrabble as a service for HOTSPOT, GRAALVM & J9.121	121
6.6	Power consumption of Dotty as a service for HOTSPOT, GRAALVM & J9. .	122
6.7	Active threads evolution when using HOTSPOT, GRAALVM, or J9.	123

List of Tables

3.1	Description of clusters included in the study	27
3.2	STD (mJ) comparison for 3 pinning strategies	35
3.3	STD (mJ) comparison when enabling/disabling the Turbo Boost	37
3.4	STD (mJ) comparison before/after tuning the OS	40
3.5	STD (mJ) comparison with/without the security patch	41
3.6	STD (mJ) comparison of experiments from 4 clusters	42
3.7	Experimental Guidelines for Energy Variations	43
4.1	the number of framework passed per test	62
4.2	Stress levels for each scenario	66
5.1	Comparison of CLBG execution times (in seconds) depending on programming languages.	81
5.2	Classification of Python implementations	99
5.3	Classification of Python implementations	101
5.4	Testing Machine Configuration	101
5.5	Energy consumption python environment	105
6.1	List of selected JVM distributions.	115
6.2	List of selected open-source Java benchmarks taken from DACAPO and RENAISSANCE.	116
6.3	Power per request for HOTSPOT, GRAALVM & J9.	121
6.4	Energy consumption when tuning JIT settings on HOTSPOT, GRAALVM & J9	126
6.5	The different J9 GC policies	127
6.6	The different HOTSPOT/GRAALVM GC policies	128
6.7	Energy consumption when tuning GC settings on HOTSPOT, GRAALVM & J9	129
6.8	J-Referral recommendations.	132

Chapter 1

Motivation

Nowadays, computers are invading our daily life from work to leisure, from fancy smartphone to an embedded peacemaker that regulates the heartbeat of people. As human beings, we are known to use tools to enhance our bodies. And thanks to computers we pushed that step even further where now we are using machines to extend our brains. Some even argue that one day they will replace us, and therefore we created our own ending. Well this is the problem for the future generations. For the moment, the main concern of humanity is to keep this planet liveable until we find another alternative.

Unfortunately those machines doesn't help that much. as according to

TODO : add the impact of the ict ..etc

Researchers are trying to solve this issue through different angles. While some scientists are trying to find an alternative green source to generate energy. others focus on reducing it'sconsumption. In computer science we are concerned with the ladder solution Therefore most of the works are done on tunning the software and the hardware in order to have a more efficient way use this energy. In this thesis we are trying to find a way to make the energy consumption of the computer to be as low as possible. Our approach is to reduce the energy consumption of the software services by changing certain parameters, such as platform, programming language, and/or the design pattern.

The best way to do so is to formulate a theory behind the energy consumption of algorithm, such as the complexity and the o notation. Unfortunately this is not possible in the current state of the art. due to the lack of knowledge about the energy consumption of the algorithms, and the strong correlation between this consumption and the hardware configuration.

Unlike algorithm optimization in the field of performance, which is agnostic toward the platform, the energy consumption of the algorithms is dependent on execution environment. Therefore, for the moment we will start by formulating some hypothesis and explore them using empirical analysis.

Our work will be presented through the following chapters:

1. 1.3: Where we discuss the work done on the energy consumption and optimization in software engineering
2. 3: It will present a set of guidelines and tools to help practitioners measure the energy consumption of their algorithms.
3. : we will present the impact of programming languages on the energy consumption of the algorithms.
4. : it will discuss the behaviour of python and the possible ways to tune it in order to reduce the energy consumption
5. : will present a study on java programming language and the impact of the JVM choice on the energy consumption
6. : as a perspective we introduce the impact of parallelism on the energy consumption on time agnostic cases

....

1.1 Motivation

....

1.2 Research Contributions

1. Introduces
2. Shows how
3. Proposes ...

1.3 Publications

1. ...
2. ...
3. ...

Chapter 2

State of the Art

Efficiency in energy usage is a well-known topic. In the majority of fields, the purpose is to minimize the energy consumption of electrical devices/components. Modern times even see energy classification (A, B,... F) for Homes, cars and electronic products, to provide the consumer an indication of the energy consumption of his devices, which will reflect on his power bill. This criterion is extended even to the hardware components of computer. Figure 2.1¹

In computer science, the objective is essentially same. Numerous studies have been conducted on energy optimization. Some of these studies concentrate on minimizing energy consumption at the hardware level, while others optimize energy consumption via software.

As an example ? evaluated the development of power use effectiveness (PUE) in data centers thkat belongs to various organizations participating in the European code of conduct for energy efficiency program [?]. The research found a gradual decline in the PUE of data centers, which measures the ratio of the overall energy supplied to the energy used by IT equipment. A low PUE implies that the majority of energy is utilized to power the data center's IT equipment, while just a little amount is needed for cooling and lighting.

we will place a greater emphasis on the software level in order to decrease the amount of energy that is used.more particularly on the execution phase of the program cycle. We will be proceeding through an empirical analysis of the energy consumption of the software while changing some components of the source code without impacting its behavior. In order to do this, we will elaborate a benchmarking process and a set of tools that are intended to assist practitioners in better comprehending and optimizing the energy usage of their applications. Thus, we will begin by examining the state of the art of empirical analysis and retrieving the best empirical experimentation methodologies in the research field.Then, we

¹<https://www.cpubenchmark.net/compare/Intel-i9-12900KS-vs-Intel-i9-12900KF/4813vs4611>

	Intel Core i9-12900KS	Intel Core i9-12900KF
Max TDP	150W	241W
Power consumption per day (kWh)	0.3	0.5
Running cost per day	\$0.075	\$0.120
Power consumption per year (kWh)	109.5	175.9
Running cost per year	\$27.38	\$43.98

Shown CPU power usage is based on linear interpolation of Max TDP (i.e. max load). Actual CPU power profile may vary.

Figure 2.1: electrical cost comparasion between two cpus

will narrow these practices down to computer science so that we can finally adapt them to energy consumption.

Section 2.1 will discuss the pitfalls and best practices associated with empirical research before applying them to our field of interest. After that Section ?? describes software energy measurements. It provides examples of hardware and software measuring instruments and describes their differences, benefits, and drawbacks. It also examines the sources of energy measurement variations, which represent a significant obstacle to achieving precise readings and a higher accuracy. Then, in section 2.3, we will go through some of the previous work on improving the energy consumption of softwares.

2.1 Benchmarking

This section will go through the flaws and best practices of empirical research before applying them to our topic of study.

2.1.1 Threats & Challenges

A successful benchmark must meet three criteria. First, it must be **reproducible** in order for others to imitate it. Second, the findings should be **accurate**, which implies that we should expect the same results each time we run the benchmark. Finally, it should **represent** reality. In other words, the experiment's findings should be applicable outside of the research lab as well. The aim of **representativeness** in this thesis is the manufacturing environment. As a result, the experiments should reflect what happens in production contexts.

Reproducibility

The reproducibility of experiments is often ranked as one of the most difficult difficulties faced by researchers. This issue has been one of the major tools science has used to help establish the validity and importance of scientific findings since the Philosophical Transactions of the Royal Society were established in 1665 [55]. In point of fact, many of the outcomes cannot be duplicated,² which led to a *replication crisis*. As this crisis hit most of the empirical studies, most of the reviews now include reproducibility as one of the minimal standard for judging scientific merit.[105] One of the criterions supporting reproducibility is the publication of the dataset and the algorithms run on the raw data to conclude the results. There is even some disagreement about what the terms "reproducibility" or "replicability" by themselves mean [48]. According to [39], *replicability* extends *reproducibility* with the ability to collect a new raw dataset comparable to the original one by re-executing the experiment under similar condition, instead of just the ability to get the same results by running the statistical analyses on the original data set.

Accuracy

According to Oxford, *accuracy* means "technical The degree to which the result of a measurement, calculation, or specification conforms to the correct value or a standard". In my case, this means the ability to run the benchmark multiple times with a low variation.

²Trouble at the lab, The Economist, 19 October 2013; [www.economist.com/news/briefing/ 21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble](http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble).

Representativeness

As obvious as it seems, the reason of executing benchmarks is to validate ideas so we can use them in the real life. However, this means that those benchmarks have to represent reality somehow. Basically, when we aim to benchmark something, we create a mock up version of the situation that we want it to work. First, we can talk about the benchmarking and their selection, then we will talk about the stress benchmark for some applications, and finally we will bring this representativeness in our case and how can we get closer to the energy consumption behavior of a software between the benchmark environment and the production one. Recently, the research community has been investigating typical "crimes" in systems benchmarking and established guidelines for conducting robust and reproducible evaluations [118].

Impact of these Challenges in the Empirical Research

In their work [119], Van-der-Kouwe *et al.* investigated 50 papers published in top venues to find out that Tier-1 papers commit an average of five benchmarking crimes. To analyze the magnitude of the phenomenon, they have identified a set of 22 "benchmarking crimes" that threaten the validity of the system.

2.1.2 Proposed Solutions

In the book "Measuring computer performance: a practitioner's guide" [78], Lilja examines performance indicators and gives in-depth treatment of benchmark program tactics. He provides clear explanations of the basic statistical methods required for interpreting measured performance data. He also outlines the overall 'design of experiments' method and demonstrates how to collect the most information with the least amount of work. This practical book will appeal to anybody seeking a comprehensive, yet intuitive, grasp of computer system performance analysis. Mytkowicz et al. made an evaluation of 133 studies from ASPLOS, PACT, PLDI, and CGO, to find out that none of the experimental findings papers appropriately considered measurement bias. Which can lead to incorrect results from an experiment if a seemingly insignificant feature of the experimental design is altered. They treated this problem by proposing two strategies for detecting measurement bias by using causal analysis and preventing it with setup randomization [92].

Bukh wrote a book about computer performance analysis, where he discussed some familiar topics that are relevant to statistical analysis, such as null hypotheses, chi squared tests, regression, discrete event simulation, Bayes' theorem, how and when to use them for experimental design, measurement, simulation, and modeling for computer systems.

the article [18] provides an intellectual framework for understanding the pervasiveness of mistake in the scientific discovery process and presents methodological, cultural, and system-level techniques for minimizing the frequency of often seen errors.

another article [104] expands on this line of thought by evaluating the uncertainty caused by replications in new research. They offer strategies for capturing uncertainty in inferential investigations, such as cross-study validation and ensemble models.

In their paper[116], the authors find that, even though it's a big move in the right direction, journal policies that require authors to give back digital scholarly objects after publication, like the data and code that back up the claims, don't get more than half of these objects back. Then, using these artifacts, about a fourth of the published computational claims in the study could be made. They suggest putting out the claims in the literature and the digital scholarly objects that back them up at the same time.

Another work, to make unify the benchmarking methodology across different research works in the field of computer science, we can cite the paper [14]. In their approach they divided any empirical experiment into four components:

1. *measurement contexts* that indicate the software and hardware components that will alter or remain constant during the experiment.
2. *workloads* which identify the benchmarks to use in the experiment, as well as their inputs;
3. *metrics* that specify the attributes to be measured and how to assess them.
4. *Data analysis* show how to examine data and evaluate the outcomes of the analysis to offer insight into the assertions that arise from the study.

Figure ?? shows the four component presented in the article [14]

2.2 Energy Measurement

To be able to reduce the energy cost of running programs we should first be able to estimate this energy consumption. Many studies have been conducted to estimate a such energy consumption. that varies from static analysis of the source code to infer its energy consumption like Pereira et al. where they provided a tool to highlight the most energy consuming parts of the code [106]. The main advantage of this approach is to be able to estimate the energy consumption of a program without running it. Unlike the complexity of programs, the energy consumption is highly related the execution environment. Therefore, static analysis might not

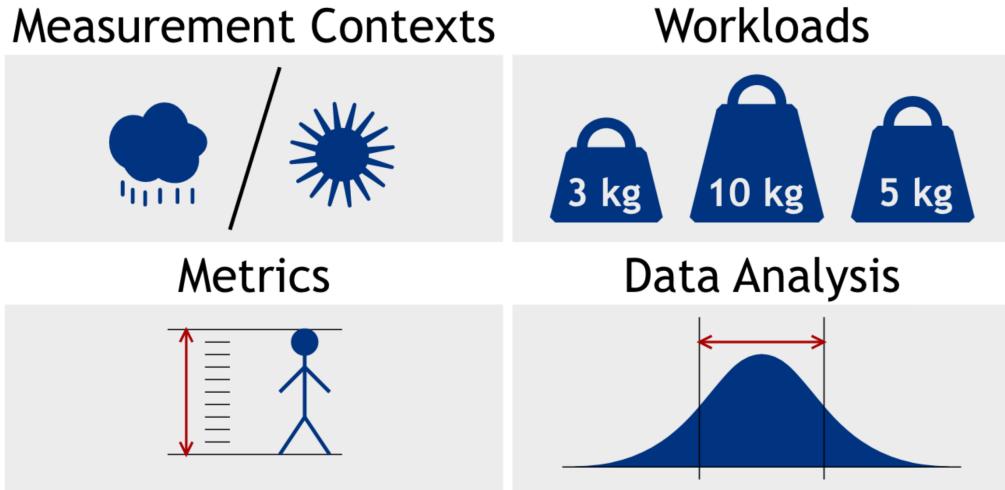


Figure 2.2: The proposed decomposition of an experiment [14]

represent the real behavior of the same program when it is run in the production environment. To treat the problem of representativeness, many researchers tend to measure the energy consumption of the programs while they are running them. Hence we will get more accurate results. there is a variety of tools to measure such energy, and they cover a large spectrum of usage depends on how **accurate** and **precise** those results are needed to be in tone hand, and the price that practitioners are ready to pay for a such accuracy/precision. Below we will present some tools that are well known in the literature and discuss their main advantages / drawbacks for our case.

2.2.1 Hardware Tools

according to Hackenberg et al. there are four main criterions to evaluate an energy measurement tool [52]

- *spatial granularity: the more specific the target of monitoring we are able to measure, the more efficient we can do optimization since we will know what causes the pitfalls of the energy consumption.*
- temporal granularity: same as spatial granularity, temporal granularity helps us to identify the sequence of code that need to be optimized.
- *scalability: this is mainly related to the cost of the tools and the ease of their integration for our system.*
- accuracy: to eliminate extra hazards and get more representative measurement.

We believe that accuracy can be extracted from those criterions since it is a result of the combination of the two first ones. therefore, we will focus more on the first 3 criterions from later on. In the following section, we will discuss some of the well known tools that are used in literature.

Nowadays, most of the high performance computing systems (HPC) implements a tool to report the energy consumption of the nodes for the sake of monitoring and administration. Those tools are mainly integrated withing the power supply units (PSU) or the power distribution units (PDU). Then, they provide an interface and a log to follow the history of the energy consumption. Despite their scalability and ease of integration. Such tools lack both in spatial and temporal granularity since they monitor the whole energy of the nodes, and most of the times they have a very low sampling frequency. Most of those tools are provided directly by the manufacturers. Such as *IBM EnergyScale technology [85, 23, Caldeira et al.]* or Dell poweredge [80], MEGware Cluststafe [17]. As said earlier, the true purpose of those tools is more monitoring than analyzing the energy consumption. WattsUp Pro, is a device that can be installed between the power source of the machine and the system under test. It allows a sampling frequency up to 1 Hz and has a internal memory to store a wide variety of data, such as the maximum voltage, current, that later can be exported via USB port for personal usage or lined to some graph programs like Logger Pro or LabQuest. The main advantage of this tool is the ability to monitor the energy consumption from a different device which will reduce the risk of interference with the energy consumption of the test [59]

despite his high temporal granularity, wattsup pro lacks in term of spatial granularity since it monitors the energy consumption of the whole system. To have a finer granularity we need to isolate the energy consumption of each component. For this, we will need more intrusive tools such.

PowerMon and its upgraded version powerMon2 [11] are based on an micro-controller chip that can monitor up to 6 channel (8 for powermon2), simultaneously. Therefore, we are able to monitor the power consumption of 4 devices at the same time. The frequency sample of this tool is up to 50 Hz, with an accuracy of 1.2%. Powermore2 comes with smaller size that can fit within 3.5 inches rack drive.

PowerInsight [73] is another fine-grained measurement tool that is based on an ARM bagelbone processor [30], which is able to measure up to 30 channel simultaneously with a frequency of 1KHz per channel.

On the other hand, GreenMiner [?], powerpack [47] in the other hand is an api that synchronizes the data gathered from other monitoring tools such as Watt's Up Pro, NI and RadioShack pro and the lines of code.

Other monitor tools have been realized by the manufactures, such as IBM Power executive [70], which allows their customer to monitor the power consumption and thermal behavior of the of BladeCenter systems in the data center.

as we see in The CPU is the part responsible of the most energy consumption of softwares. hence the finer we go to measure this energy consumption the better it is for our work. Fortunately, Intel and later AMD proposed a tool that estimate the power consumption of different part of the CPI based on counter performances. RAPL (*RUnning Average Power Limit*) [51, 53] is a set of registers that was introduced by Intel in their CPU since Sandy bridge generation, and later it was followed by AMD, since Family 17h Zen. It is capable of monitoring different components of CPU, such as DRAM, cores, and the integrated GPU for desktop processors.

The advantage of such an approach is that the absence of any intrusive measurement tools. Furthermore, they have a high temporal granularity with a sampling frequency up to 1 KHz [60].

With similar approach, we can find NVIDIA reporting tools, such as GPU TESLA [21] and the NVML library [43].

2.2.2 Software Tools

Software-based measurement tools are based on other hardware tools to monitor energy use. Granularity is the core value of these technologies, unlike hardware tools, which only provide the total energy usage of the system/component (computer, server, motherboard, etc.) for most cases. Because they are frequently constructed on empirical estimations and data learning methods, they drop in accuracy.

Many software measuring tools understand the behavior of a power model and provide estimates of energy usage. This model is then used to allocate the observed energy consumption among different execution entities, such as processes, control groups, threads, or even code lines.

The first examples of software measurement tools are PowerAPI [31], SmartWatts formula [44] and SelfWatts [45]. These tools collect global energy consumption measurements from RAPL and use other system events such as cache misses/hits and CPU frequency evolution (DVFS) via a sensor to construct a power model of the control groups (system control groups, docker containers, kubernetes pods, etc.) using a Ridge regression. The model continuously learns and improves its precision of real-time energy usage data with a maximum frequency of 100 Hz. The instrument has a decentralized, lightweight design. Only the lightest sensors required for data collection and transmission are put into the monitored devices. The SmartWatts formula is then executed on the primary server in order to

construct the model that permits assigning the energy usage for each operating control group. PowerAPI is only compatible with Linux on a bare-metal physical computer.

WattWatcher [74] is a multi-core power measuring framework that provides process-level energy measurements. This program uses power models to predict process energy usage. It uses CPU events which are passed from the measured node to a model generator node in order to construct the power model. It works by combining a description of the CPU with a list of the hardware events through multiple calibration phases in order to build a robust model.

Joulemeter [71, 63] is a Microsoft software that estimates the energy usage of Windows running applications down to the process level by using power models (for CPU, memory, and drives). It employs low-overhead power models to infer power consumption from resource utilization during runtime, and it provides power limiting features for virtual machines. Previous Joulemeter tests [64] demonstrated that the instrument provides a less accurate estimation of energy use that differs greatly from the real one. To adjust its models to the hardware on which it operates, Joulemeter must first go through a calibration step. It is able to only monitor one process at a time with a frequency of 1 Hz.

jRAPL is another example of an energy measurement tool estimating tool that has been utilized in a variety of publications[50?]. This software enables for the energy usage of Java programs, functions, or even a block of code lines to be profiled and measured. The measurements are heavily reliant on the data supplied by RAPL. As a result, the global energy consumption collected by RAPL between two timestamps (the start and finish of the code to measure) is used to calculate the energy consumption of the Java code. Tests using jRAPL should be done on a well-configured machine to minimize the impact of the operating system and user operations on the overall energy consumption of jRAPL.

Another process-level energy usage measuring tool is Jolinar [62, 96]. The tool does not need a calibration phase and relies on pre-established power models based on hardware metrics (TDP, disk I/O rate, and so on). These settings must be identified and supplied by the user for his machine. Jolinar is only capable of measuring the energy consumption of one application at a time. At the end of the execution, the tool provides the CPU, DRAM, and disk energy usage of the main process.

Jalen [97] is another tool that profiles and monitors the energy usage of a Java program. Unlike jRAPL, Jalen is able to cover the scope down to the function level. It gathers data using code instrumentation and statistical sampling at a predetermined pace. Because of the overhead that code instrumentation may incur, the authors recommend utilizing the second option. Every 10 ms, Jalen records the JVM's stack trace together with the CPU time of

threads and computes statistics about method calls. These statistics are then utilized to calculate each method's energy usage.

selfwatts [45] powerjoular [94]

2.2.3 Hybrid Tools

Simgrid simulation of power consumption of parallel application using simgrid [?]
impact of manufacturing process on the energy variation [?]
different energy consumption between identical processors [idle and hight load][123]
the temperature is the main raison a high energy variation [?]
external factors are impact on the energy variation [90]
the place of the server in the room doesn't affect the energy variation [?]
comparaison of three measuring tools and they did exhebit 10% variation each time [?]
low-cost scalable variation-aware algorithm [?]
reducing the energy variation by disabling turbo boost [?]
reducing the energy consumption in parallel systems [?]
[?]

2.3 Programming Languages & Performances

2.3.1 Energy

energy comparaison of programming languages in the game benchmark by debian [?]
toward green ranking [34]
java vs kotlin in web performances [19]
rosetta code [93] [86]
network energy comparaison [7]
Aequitas [113]
vm placement [88]
Mishra et al.
static cost of Vms [72]
carbon footprint of training neural network [?]
SPELL, the energy leaks detector tool [?]
impact of energy profiling on software [63]
impact of maintainability on software energy evolution [24]
definition [124]
comparison [26]

empirical study [36]
impact of website implementation on energy consumption [108] [?]
PHP [13] [?]
simulation of web cache [25]
java spring analysis [46]
performance [87]
cross compiler benchmarking [127]

2.3.2 Python

web frameworks on python [Pankiv]

hope library [4]
bytecode effect [12]
numba [35]
intel python [76]
python 2 vs 3 [89]
python runtime performances [112] [91]
static vs dynamic languages [Pang et al.]
concurrent benchmark framework [?]

2.3.3 JVM

java vs python [38]

hotspot vs J9 [98] - same but for big queries [28]
constant overheating of the jvm [?]
infer the energy cost based on the byte code instructions [81]
java collection energy footprint [?] [?]
java classes energy footprint [?]
framework to reduce java collections [?]
energy consumption of java dev tools [10]
Microbenchmarks on jvm [?] [10]
android automatic refactoring to reduce energy consumption [9] [?]
java vs native c in android [33]

2.3.4 benchmarking

NAS [Bailey et al.] statistics [57] benchmark crimes [119] microservices [49] impact of webservers on web applications energy [82]

2.4 Energy Variation

In theory, using identical CPU, same memory configuration, similar storage and networking capabilities, should increase the accuracy of physical measurements. Unfortunately, this is not possible when it comes to measuring the energy consumption of a system. Applying the benchmarking guidelines and repeating the same experiment with in the same configuration are not sufficient to reproduce the the same energy measurements, not only between identical machines, but even within the same machine. This difference—also called *energy variation* (EV)—has become a serious threat to the accuracy of experimental evaluations.

Figure 2.3 illustrates this variation problem as a violin plot of 20 executions of the benchmark *Conjugate Gradient* (CG) taken from the *NAS Parallel Benchmarks* (NBP) suite [6], on 4 nodes of an homogeneous cluster (the cluster Dahu described in Table 3.1) at 50 % workload. We can observe a large variation of the energy consumption, not only among homogeneous nodes, but also at the scale of a single node, reaching up to 25 % in this example.

Some researchers started investigating the hardware impact of the energy variation of power consumption. As an example, one can cite [16, 117] who reported that the main cause of the variation of the power consumption between different machines is due to the **CMOS** manufacturing process of transistors in a chip. [58] described this variation as a set of parameters, such as CPU Frequency and the thermal effect.

2.4.1 Studying Hardware Factors

This variation has often been related to the manufacturing process [29], but has also been a subject of many studies, considering several aspects that could impact and vary the energy consumption across executions and on different chips. On the one hand, the correlation between the processor temperature and the energy consumption was one of the most explored paths. Kistowski *et al.* showed in [65] that identical processors can exhibit significant energy consumption variation with no close correlation with the processor temperature and performance. On the other hand, the authors of [125] claimed that the processor thermal effect is one of the most contributing factors to the energy variation, and the CPU temperature and the energy consumption variation are tightly coupled.

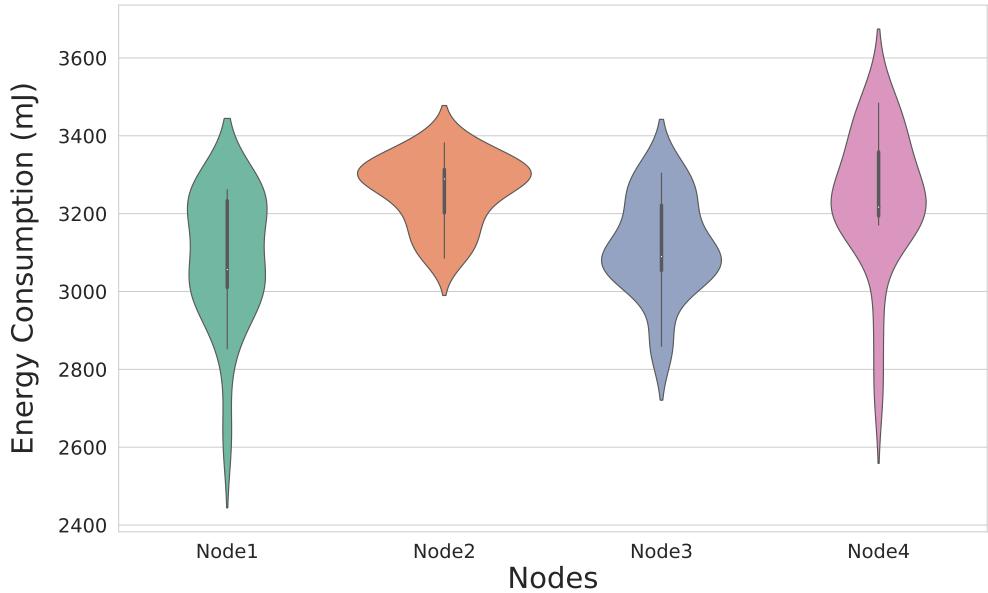


Figure 2.3: CPU energy variation for the benchmark CG

TODO : add the corelation value

This exposes the processor temperature as a delicate factor to consider while comparing energy consumption variations across a set of homogeneous processors.

The ambient temperature was also discussed in many papers as a candidate factor for the energy variation of a processor. In [121], the authors claimed that energy consumption may vary due to fluctuations caused by the external environment. These fluctuations may alter the processor temperature and its energy consumption. However, the temperature inside a data center does not show major variations from one node to another. In [42], El Mehdi Dirouri *et al.* showed that switching the spot of two servers does not affect their energy consumption. Moreover, changing hardware components, such as the hard drive, the memory or even the power supply, does not affect the energy variation of a node, making it mainly related to the processor. This result was recently assessed by [125], where the rack placement and power supply introduced a maximum of 2.8 % variation in the observed energy consumption.

Beyond hardware components, the accuracy of power meters has also been questioned. Inadomi *et al.* [61] used three different power measurement tools: RAPL, Power Insight³

³<https://www.itssolution.com/products/trellis-power-insight-application>

and BGQ EMON. All of the three tools recorded the same 10 % of energy variation, that was supposedly related to the manufacturing process.

2.4.2 Mitigating Energy Variations.

Acknowledging the energy variation problem on processors, some papers proposed contributions to reduce and mitigate this variation. In [61], the authors introduced a variation-aware algorithm that improves application performance under a power constraint by determining module-level (individual processor and associated DRAM) power allocation, with up to $5.4 \times$ speedup. The authors of [54] proposed parallel algorithms that tolerate the variability and the non-uniformity by decoupling per process communication over the available CPU. Acun *et al.* [3] found out a way to reduce the energy variation on Ivy Bridge and Sandy Bridge processors, by disabling the Turbo Boost feature to stabilize the execution time over a set of processors. They also proposed some guidelines to reduce this variation by replacing the old—slower—chips, by load balancing the workload on the CPU cores and leaving one core idle. They claimed that the variation between the processor cores is insignificant. In [27], the researchers showed how a parallel system can be used to deal with the energy variation by compensating the uneven effects of power capping.

In [83], the authors highlight the increase of energy variation across the latest Intel micro-architectures by a factor of 4 from Sandy Bridge to Broadwell, a 15 % of run-to-run variation within the same processor and the increase of the inter-cores variation from 2.5 % to 5 % due to hardware-enforced constraints, concluding with some recommendations for Broadwell usage, such as running one hyper-thread per core.

2.5 Conclusion

As we have seen in the state of the art, many methods have been proposed to reduce the energy footprint of the the ICT, which they can be applied in different parts of the lifecycle of the program, from consumption to the execution. Furthermore, the execution phase took attention of many researchers because it is the part where the most energy is consumed. This thesis will focus on that aspect as well. However, unlike the most work that have been done on the hardware aspect, we will target the software impact on this energy, starting from the choice of the programming language up to how to tune some features of a frameworks in order to make the software consumes less energy. To do so we use the empirical approach due to its consistency for the moment. Unlike the performance, which is essentially related to

the complexity of the algorithm, the energy consumption is more impacted by the hardware. Therefore, to optimize the energy consumption we choose a spiral method based on 3 phases:

1. execute the code,
2. measure the program,
3. infer the guidelines.

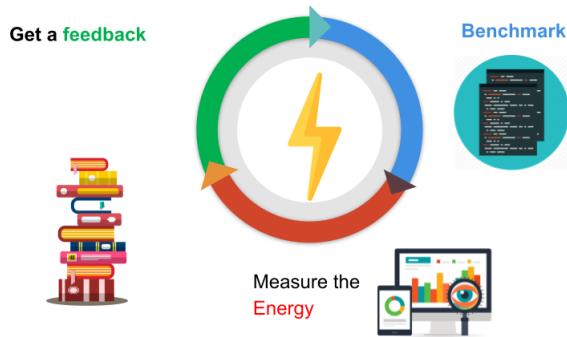


Figure 2.4: the spiral method of energy optimization

The aim of this work is to present a set of guidelines to create a benchmarking system to measure the energy consumption of different programs. After that, we will use this system to compare the energy consumption of different programming languages. We will extend the work of ? to a closer distance to production environment by comparing a set of use cases. Starting by GRPC framework, and a set of Web Frameworks. Finally, we will discuss the impact of the execution environment on the energy consumption of two of the most famous programming languages Java and Python and present how can tuning the Virtual Machine can reduce the energy consumption.

Part I

Benchmarking Programming Technologies

Chapter 3

A Benchmarking Protocol to Measure Software Energy Consumption

3.1 Motivation

Empirical measurements are critical to capture the effect of developers' choices on software energy consumption. To accomplish this, one should not overlook the benchmarking pitfalls highlighted in the state of the art [119]. Furthermore, when it comes to evaluating the energy consumption of software, other challenges occur, such as the impact of hardware, the difference between the machines used for measurements and those that may apply the results, and so on. Moreover, considerable advances in computer science have resulted in an increase in the number of obsolete results. Addressing all of them becomes significantly more difficult in studies that compare multiple treatments.

Furthermore, between the exploratory experiment and the published results, there may have been the appearance of new candidates as well as the evolution of others. As a result, it is not only necessary to ensure the "reproducibility" of the results, but it is also essential to provide room for new candidates. As a result, we'd like to present "Extension", a new approach for leveraging experimentation. The ability to give not only the tools needed to replicate the tests but also to integrate more candidates, workloads, or important key performance indicators This chapter covers ways to overcome empirical analysis challenges in energy consumption studies. First, we go through the three components of a successful benchmark within the energy consumption experiment. Section 3.2 concentrates on the "reproducibility" issue and how to deliver reproducible experiments without interfering with the energy measurement. Section 3.3 discusses accuracy in software energy consumption. Finally, Section 3.3.7 presents tools for introducing new candidates, workloads, or other

key performance indicators. Regarding the "representativeness" of the benchmarks, we will elaborate in the section entitled Perspectives.

3.2 Reproducibility within the context of energy

This section discusses various techniques to encapsulate the systems-under-test in order to assure the reproducibility of an experiment.

Virtual Machines The very first choice should be to use *Virtual Machines* (VM), which allow researchers to select the most appropriate tools, software, and operating system that they are most comfortable with without incurring the cost of changing the actual working environment, giving them more control over dependencies and the execution environment. Furthermore, adopting a VM addresses the *replication crisis* since virtual images allow even the most sophisticated architecture to be simply replicated by instantiating a copy of the image.

This option, however, comes at a cost. Because of the hypervisor, software will be built on two kernels: one for the virtual machine (guest) and one for the host machine, resulting in a visible overhead and a negative influence on the performance of the system-under-test. As a result, we cannot use VM for performance-related tests. Isolation is another drawback of VM: while this feature protects the experimental setting from unwanted interference from the outside world, it is possible that this interaction is required—especially if the experiment is dependent on an external source such as energy monitors.

Containers Another option would be to use something that allows us to benefit from the host OS's isolation while simultaneously simplifying replication as proposed by VM and the direct interface with the hardware provided by the traditional techniques.

Containers provide such an advantage by ensuring application separation and ease of replication. Figure 3.1 depicts the architectural differences between virtualization and container technologies. There are three main types of virtualization.

- Type 1: runs on the hardware directly. It is primarily utilized by cloud providers where there is no host OS and only VMs that run on the open-source Xen or VMware ESX hypervisors.
- Type 2: runs on top of the host operating system and is most commonly found on personal computers. VMware server and virtualBox are notable examples of this

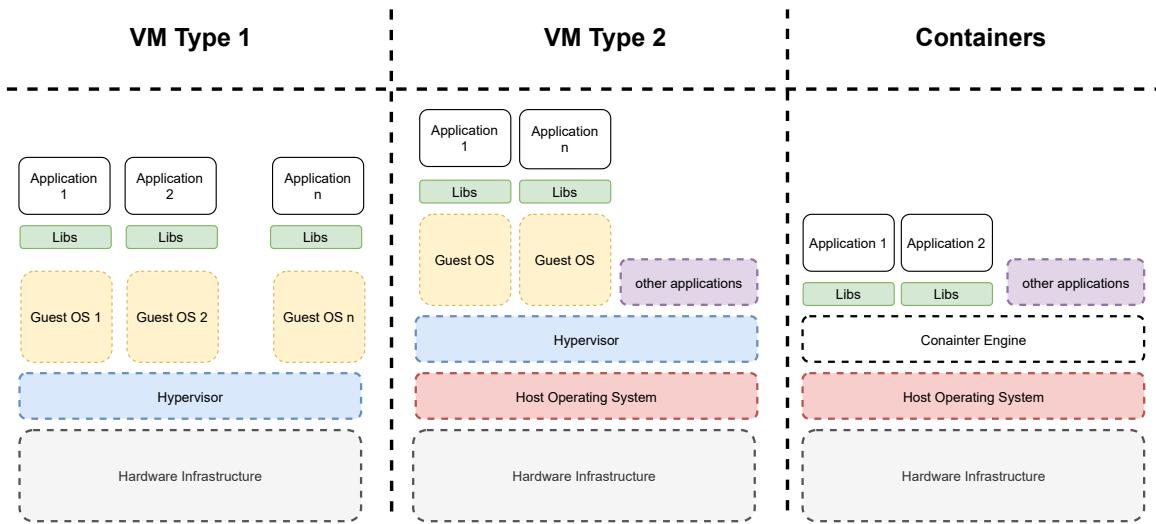


Figure 3.1: Different Methods of Virtualization

category, and most researchers' experiments use them. However, because of the two operating systems, the applications are typically slower.

- **Containers:** run their operating systems on the host kernel rather than their own , which makes them smaller, faster, and more efficient in terms of hardware utilization. One can cite *Docker*, *Linux LXC*, or *LXD* [2].

3.2.1 Docker Vs. Virtual Machine

Despite the fact that Type 1 is more performant than Type 2, the latter is the most used in research, as most researchers tend to conduct their experiments on their own machines. Docker, on the other hand, is the most well-known container technology. In our case, we are more likely to promote Docker for two reasons:

1. As previously stated [143] [reference Morabito (2015) and van Kessel et al. (2016)], we require a lightweight orchestrator to limit the overhead on energy usage of our experiments [120] [cite Morabito (2015) and van Kessel et al. (2016)],
2. We need to communicate with the host OS because we are using hardware sensors to measure the energy consumption.

3.2.2 Docker & Energy

Now that we have decided to use container technology to enclose our tests, what effect will this have on the amount of energy our tests use?

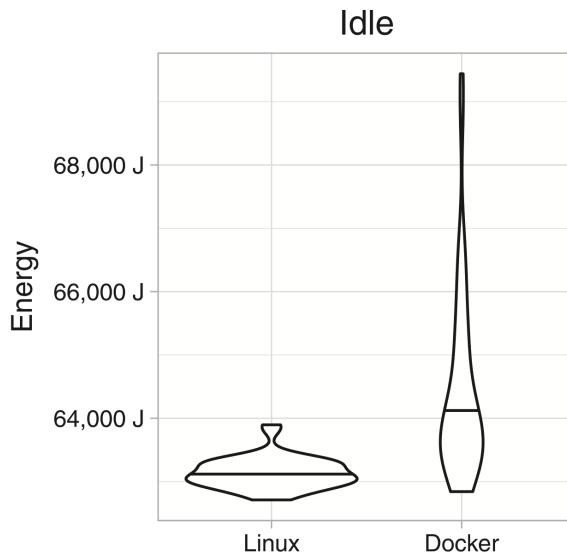


Figure 3.2: energy consumption of Idle system with and without docker [?]

Using research from [?] who examined how adding the Docker layer affected energy consumption. ? conducted their experiment by running numerous benchmarks both with and without Docker. They contrasted the energy usage and execution time that resulted. The first step was to observe the effects of the orchestrator and the Docker daemon while there was no work to be done. Then, they use three benchmarks in their experiments: Wordpress, Redis and PostgreSQL. The values below show the system under test's energy consumption while it is idle. Docker has an overhead of roughly 1,000 joules, as seen in Figure 3.2 .

Yet, as seen in Figure 3.3, Docker increased the execution duration of the benchmark by 50 seconds, which led to a significant rise in the energy usage. According to the authors, this overhead is primarily caused by the Docker daemon and not by the fact that the application is in a container.

Furthermore, they calculated the cost of this extra energy, which was less than 0.15\$ in the worst-case scenario, which is insignificant in comparison to the benefits of Docker for isolation and reproducibility.

To summarize, Docker-based software tends to consume more energy since it takes longer to execute. The execution of the Docker daemon causes an increase in average power consumption of only **2 Watts**. This overhead can reach up to 5% in IO-intensive applications, while it is barely visible in CPU- or DRAM-intensive workloads.

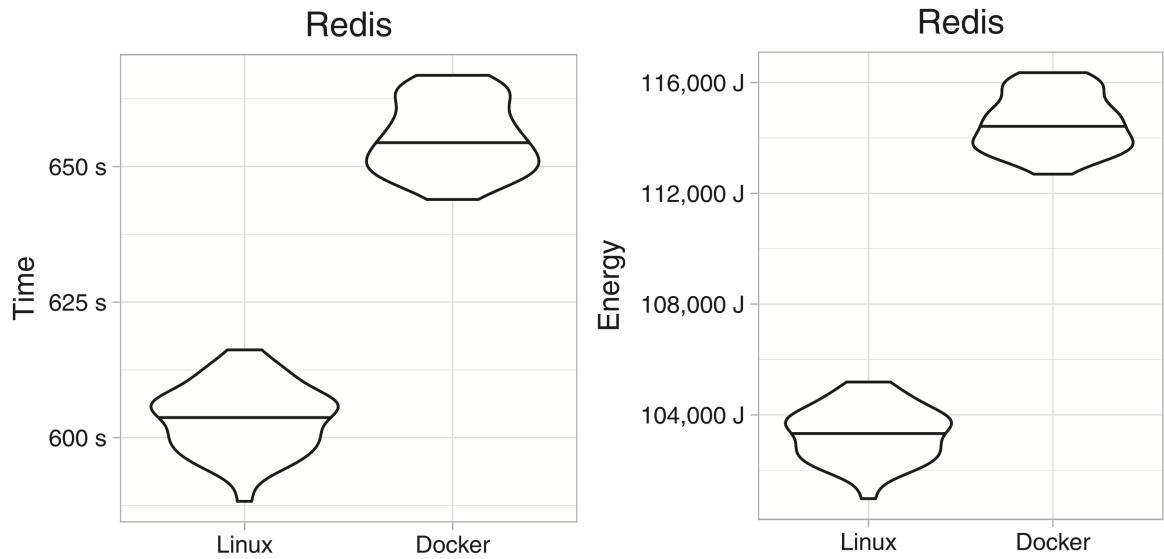


Figure 3.3: Execution time & energy consumption of Redis with and without Docker [?]

3.2.3 Conclusion

As can be seen, the introduction of the supervisor has increased the impact of Docker on energy that will be applied equitably to all experiments. Therefore when it comes to comparison analysis, it will mitigate its impact automatically. Furthermore, because we have access to the host Hardware, we do not need to worry about capturing the SUT's energy use. As a result, we will use Docker as an encapsulation solution for all of our experiments. It will ensure the reproducibility of our experiments in a simple and straightforward manner.

3.3 Improving Accuracy by Taming Energy Variations

While the previous section aimed to ensure the reproducibility of our software energy consumption experiments, this section provides a collection of tips and tools to help increase the accuracy of these experiments. We are well aware of the effect that hardware has on energy fluctuations. However, we feel that there is still an opportunity for practitioners to minimize this energy variance by employing solely tuneable factors. To that end, we conducted a series of empirical studies utilizing state-of-the-art recommendations to discover which controllable elements can limit the variations of benchmark energy usage.

In this section, we investigate the following research questions:

RQ 1: Does the benchmarking protocol affect the energy variation?

RQ 2: How important is the impact of the processor features on the energy variation?

RQ 3: What effect does the operating system have on energy variation?

RQ 4: Does the choice of processor make a difference in reducing the energy variation?

3.3.1 Experimental Setup

This section describes our detailed experimental environment, covering the cluster configuration and the benchmarks we used to justify our experimental methodology.

Measurement Context

There are three main contexts.

- different machines with different settings;
- different machines with the same settings ;
- the same machine.

We used the platform Grid5000 (G5K) [8, 84], which is a large, flexible testbed for experiment-driven research that is spread out across all of France, to meet these needs. Grid5000 has several clusters that are made up of 4 to 124 identical machines with different configurations for each cluster. For our experiment, we looked at four groups. Our main criterion was the CPU configuration. The table 3.1 below, presents a description of the four clusters that have been chosen for our experiments.

Table 3.1: Description of clusters included in the study

Cluster	Processor	Nodes	RAM
Dahu	2× Intel Xeon Gold 6130	32	192 GiB
Chetemi	2× Intel Xeon E5-2630v4	15	768 GiB
Ecotype	2× Intel Xeon E5-2630Lv4	48	128 GiB
Paranoia	2× Intel Xeon E5-2660v2	8	128 GiB

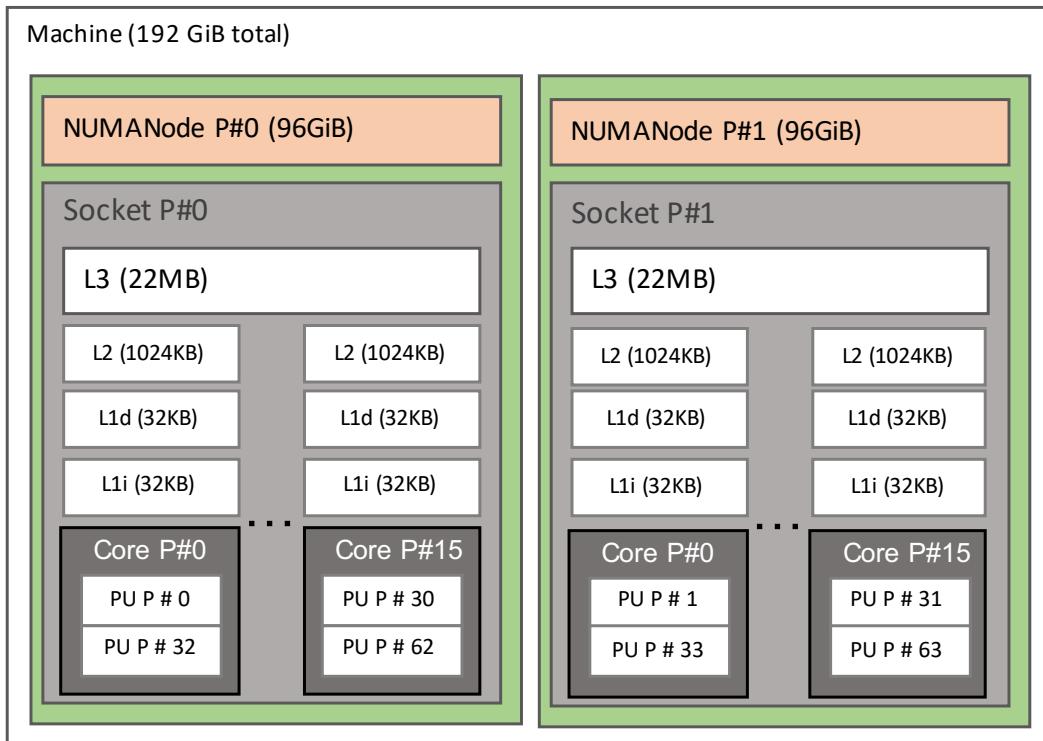


Figure 3.4: Topology of the nodes of the cluster Dahu

As most of the nodes are equipped with two sockets (physical processors), we use the acronym CPU or socket to designate one of the two sockets and PU for the *processing unit*. For our study we consider hyper-threads as distinct PUs.

Figure 3.4 shows the detailed topology of a node in the cluster Dahu as an example.

Workload

We picked the benchmarks based on two criteria.

First, **scalability**: We wanted to learn as much as possible about the experiment in the time we had, so we needed some benchmarks that could grow or shrink with the number of PUs and work in different situations. The second criterion is whether or not the workload

is **representative**. As stated in the challenges, a workload must represent the production environment or the experiment would be inconsistent [14].

To meet these criteria, we looked at the "state of the art" and found the most common benchmarks used to test the performance of hardware. We then chose the ones that could be scaled up. Our candidate is emphNAS Parallel Benchmark (NPB v3.3.1) [6]: one of the most used benchmarks for *HPC*. We used the applications (LU), the *Conjugate Gradient* (CG) and *Embarrassingly Parallel* (EP) computation-intensive benchmarks in our experiments, with the C data class. Furthermore we have used other applications to validate our results using more applications such as Stress-ng v0.10.0,¹ pbzip2 v1.1.9,² linpack³ and sha256 v8.26.⁴

Metrics & Measurement Tools

Our metric for the accuracy of the test is the Standard deviation aka **STD** of the energy consumption. Therefore whether the tests consume more or less energy is out of our scope. To study this variation we need first a tool to measure the energy consumption. For this we used POWERAPI [32], which is a power monitoring toolkit that is based on *Intel Running Average Power Limit* (RAPL) [68]. The advantage of PowerAPI is that it reports the Energy consumption of CPU and DRAM at a socket level.

Our testbeds are run with a minimal version of Debian 9 (4.9.0 kernel version)⁵ where we install Docker (version 18.09.5), which will be used to run the RAPL sensor and the benchmark itself. The energy sensor collects RAPL reports and stores them in a remote MON-GODB instance, allowing us to perform *post-mortem* analysis in a dedicated environment. Using Docker makes the deployment process easier on the one hand, and provides us with a built-in control group encapsulation of the conducted tests on the other hand. This allows POWERAPI to measure all the running containers, even the RAPL sensor consumption, as it is isolated in a container.

Every experiment is conducted on 100 iterations, on multiple nodes and using the 3 NPB benchmarks we mentioned, with a warmup phase of 10 iterations for each experiment. In most cases, we were seeking to evaluate the *STandard Deviation* (STD), which is the most representative factor of the energy variation. We tried to be very careful, while running our experiments, not to fall in the most common benchmarking "crimes" [118]. As we study the STD difference of measurements we observed from empirical experiments, we use the

¹<https://kernel.ubuntu.com/~cking/stress-ng>

²<https://launchpad.net/pbzip2/>

³<http://www.netlib.org/linpack>

⁴<https://linux.die.net/man/1/sha256sum>

⁵<https://github.com/grid5000/environments-recipes/blob/master/debian9-x64-min.yaml>

bootstrap method [41] to randomly build multiple subsets of data from the original dataset, and we draw the STD density of those sets, as illustrated in Figure 3.5. Given the space constraints, we report on aggregated results for nodes, benchmarks and workloads. However, the raw data we collected is available through the public repository we published.⁶ We believe this can help to achieve better and more reliable comparisons. We mainly consider 3 different workloads in our experiments: single process, 50 %, and 100 %, to cover the low, medium and high CPU usage when analyzing the studied parameters effect, respectively. These workloads reflect the ratio of used PU count to the total available PU.

3.3.2 Analysis

In this part, we aim to establish experimental guidelines to reduce the CPU energy variation. We therefore explore many potential factors and parameters that could have a considerable effect on the energy variation.

3.3.3 Docker & Accuracy

As the state of the art assesses the impact of Docker on the energy consumption, one can also consider its impact on accuracy. In other words:

RQ: does Docker affect the energy variation of the experiments?

To answer this question we conducted a preliminary experiment by running the same benchmarks LU, CG and EP in a Docker container and a flat binary format on 3 nodes of the cluster Dahu to assess if Docker induces an additional variation. Figure 3.5 reports that this is not the case, as the energy consumption variation does not get noticeably affected by Docker while running a same compiled version of the benchmarks at 5 %, 50 % and 100 % workloads. In fact, while Docker increases the energy consumption due to the extra layer it implements [40], it does not noticeably affect the energy variation. The *standard deviation* (STD) is even slightly smaller ($STD_{Docker} = 192mJ, STD_{Binary} = 207mJ$), taking into account the measurements errors and the OS activity.

RQ 1: Benchmarking Protocol

To achieve a robust and reproducible experiment, practitioners often tend to repeat their tests multiple times, in order to analyze the related performance indicators, such as execution time, memory consumption or energy consumption. We therefore aim to study the bench-

⁶<https://github.com/anonymous-data/Energy-Variation>

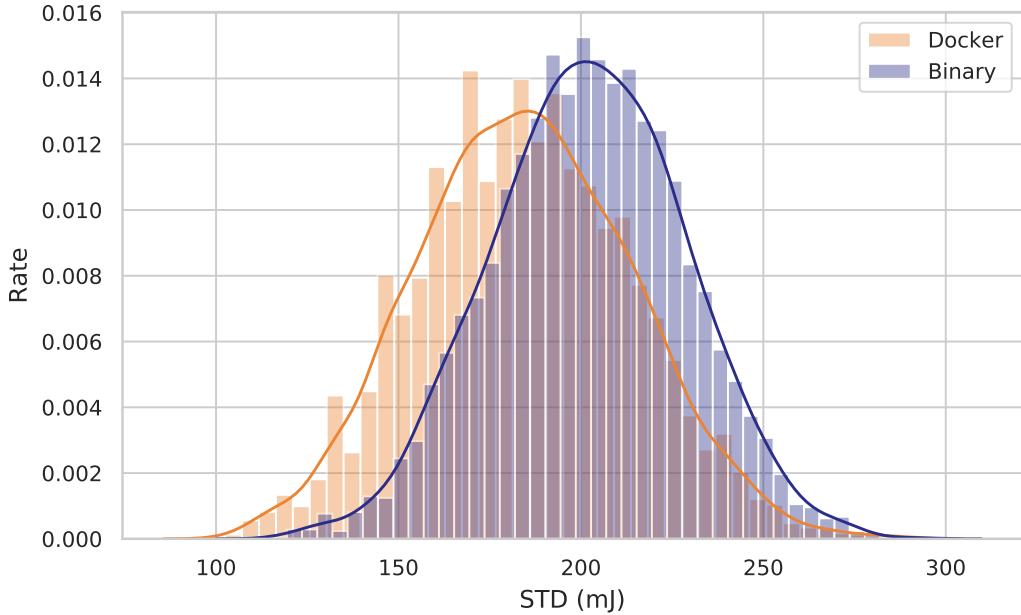


Figure 3.5: Comparing the variation of binary and Docker versions of aggregated LU, CG and EP benchmarks

marking protocol to identify how to efficiently iterate the tests to capture a trustable energy consumption evaluation.

In this first experiment, we investigate if changing the testing protocol affects the energy variation. To achieve this, we considered 3 execution modes: In the "normal" mode, we iteratively run the benchmark 100 times without any extra command, while the "sleep" mode suspends the execution script for 60 seconds between iterations. Finally, the "reboot" mode automatically reboots the machine after each iteration. The difference between the normal and sleep modes intends to highlight that the CPU needs some rest before starting another iteration, especially for an intense workload. Putting the CPU into sleep for several seconds could give it some time to reach a lower frequency state or/and reduce its temperature, which could have an impact on the energy variation. The reboot mode, on the other hand, is the most straightforward way to reset the machine state after every iteration. It could also be beneficial to reset the CPU frequency and temperature, the stored data, the cache or the CPU registries. However, the reboot task takes a considerable amount of time, so rebooting the node after every single operation is not the fastest nor the most eco-friendly solution, but it deserves to be checked to investigate if it effectively enhances the overall energy variation or not.

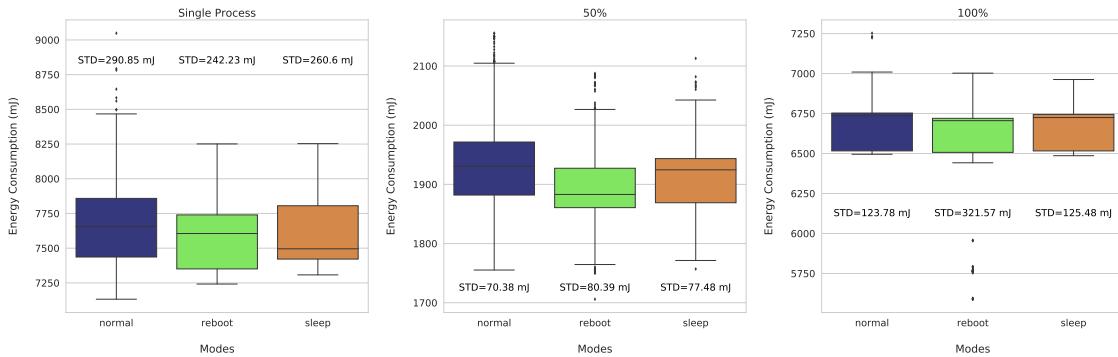


Figure 3.6: Energy variation with the normal, sleep and reboot modes

Figure 3.6 reports on 300 aggregated executions of the benchmarks LU, CG and EP, on 4 machines of the cluster Dahu (cf. Table 3.1) for different workloads. We note that the results have been executed with different datasets sizes (B, C and D for single process, 50 % and 100 % respectively) to remedy to the brief execution times at high workloads for small datasets. This justifies the scale differences of reported energy consumptions between the 3 modes in Figure 3.6. As one can observe, picking one of these strategies does not have a strong impact on the energy variation for most workloads. In fact, all the strategies seem to exhibit the same variation with all the workloads we considered—*i.e.*, the STD is tightly close between the three modes. The only exception is the reboot mode at 100 % load, where the STD is 150 % times worst, due to an important amount of outliers. This goes against our expectation, even when setting a warm-up time after reboot to stabilize the OS.

In Figure 3.7, we study the standard deviation of the three modes by constituting 5,000 random 30-iterations sets from the previous executions set and we compute the STD in each case, considering mainly the 100 % workload as the STD was 150 % higher for the reboot mode with that load. We can observe that the considerable amount of outliers in the reboot mode is not negligible, as the STD density is clearly higher than the two other modes. This makes the reboot mode as the less appropriate for the energy variation at high workloads.

To answer RQ 1, we conclude that the benchmarking protocol **partially affects** the energy variation, as highlighted by the reboot mode results for high workloads.

RQ 2: Processor Features

The C-states provide the ability to switch the CPU between more or less consuming states upon activities. Turning the C-states on or off have been subject of many discussions [122], because of its dynamic frequency mechanism but, to the best of our knowledge, there have been no fully conducted C-states behavior analysis on CPU energy variation.

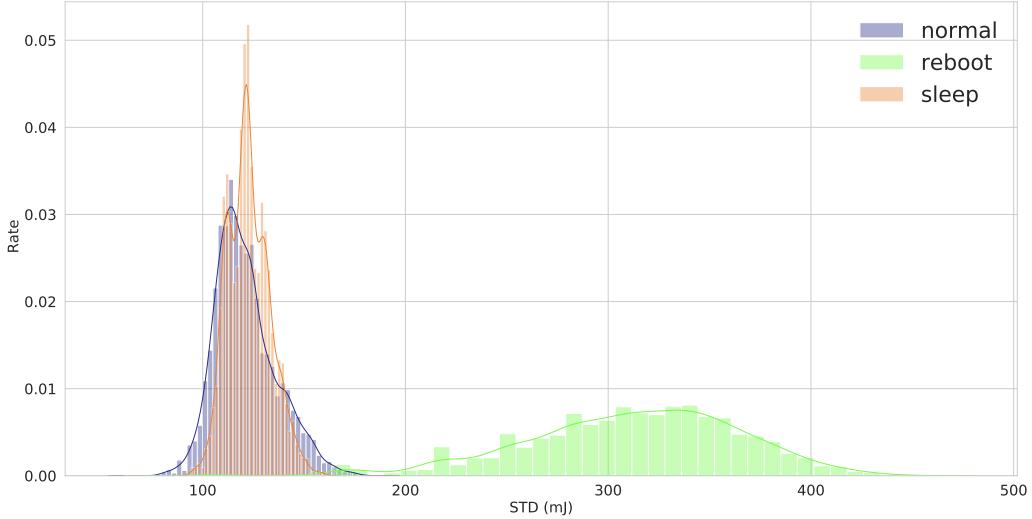


Figure 3.7: STD analysis of the normal, sleep and reboot modes

We intend to investigate how much the energy consumption varies when disabling the C-states (thus, keeping the CPU in the C0 state) and at which workload. Figure 3.8 depicts the results of the experiments we executed on three nodes of the cluster Dahu. On each node, we ran the same set of benchmarks with two modes: C-states on, which is the default mode, and C-states off. Each iteration includes 100 executions of the same benchmark at a given workload, with three workload levels. We note that our results have been confirmed with the benchmarks LU, CG and EP.

We can clearly see the effect that has the C-states off mode when running a single-process application/benchmark. The energy consumption varies 5 times less than the default mode. In this case, only one CPU core is used among 2×16 physical cores. The other cores are switched to a low consumption state when C-states are on, the switching operation causes an important energy consumption difference between the cores, and could be affected by other activities, such as the kernel activity, causing a notable energy consumption variation. On the other hand, switching off the C-states would keep all the cores—even the unused ones—at a high frequency usage. This highly reduces the variation, but causes up to 50 % of extra energy consumption in this test ($Mean_{C\text{-states-off}} = 11,665mJ, Mean_{C\text{-states-on}} = 7,641mJ$).

At a 100 % workload, disabling the C-states seems to have no effect on the total energy consumption nor its variation. In fact, all the cores are used at 100 % and the C-states module would have no effect, as the cores are not idle. The same reason would apply for the 50 % load, as the hyper-threading is active on all cores, thus causing the usage of most of them.

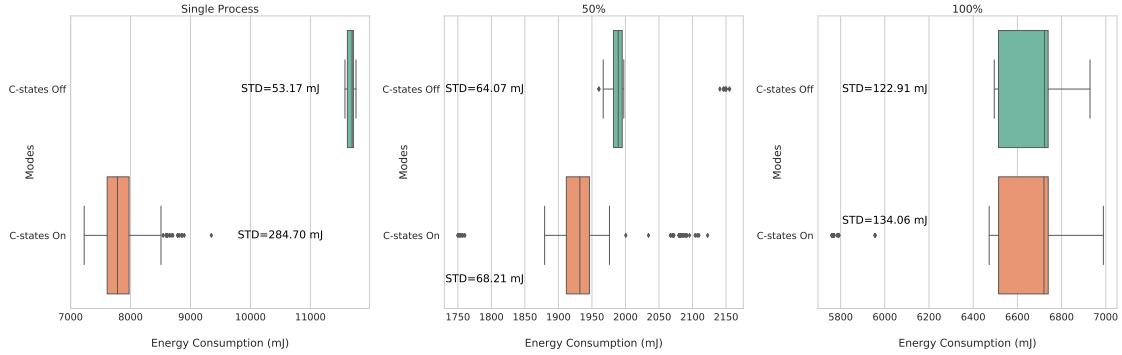


Figure 3.8: Energy variation when disabling the C-states

For single process workloads, disabling the C-states causes the process to consume 50 % more energy as reported in Figure 3.8, but reduces the variation by 5 times compared to the C-states on mode. This leads to mainly two questions: Can a process pinning method reduce/increase the energy variation? And, how does the energy consumption variation evolve at different PU usage level?

Cores Pinning To answer the first question, we repeated the previous test at 50 % workload. In this experiment, we considered three cores usage strategies, the first one (S1) would pin the processes on all the PU of one of the two sockets (including hyper-threads), so it will be used at 100 %, and leave the other CPU idle. The second strategy (S2) splits the workload on the two sockets so each CPU will handle 50 % of the load. In this strategy, we only use the core PU and not the hyper-threads PU, so every process would not share his core usage (all the cores are being used). The third strategy (S3) consists also on splitting the workload between the two sockets, but considering the usage of the hyper-threads on each core—*i.e.*, half of the cores are being used over the two CPU. Figure 3.9 reports on the energy consumption of the three strategies when running the benchmark CG on the cluster Dahu. We can notice the big difference between these three execution modes that we obtained only by changing the PU pinning method (that we acknowledged with more than 100 additional runs over more than 30 machines and with the benchmarks LU and EP). For example, S2 is the least power consuming strategy. We argue that the reason is related to the isolation of every process on a single physical core, reducing the context switch operations. In the first and third strategy, 32 processes are being scheduled on 16 physical cores using the hyper-threads PU, which will introduce more context switching, and thus more energy consumption.

We note that even if the first and third strategies are very similar (both use hyper-threads, but only on one CPU for the first and on two CPU for the third), the gap between them is

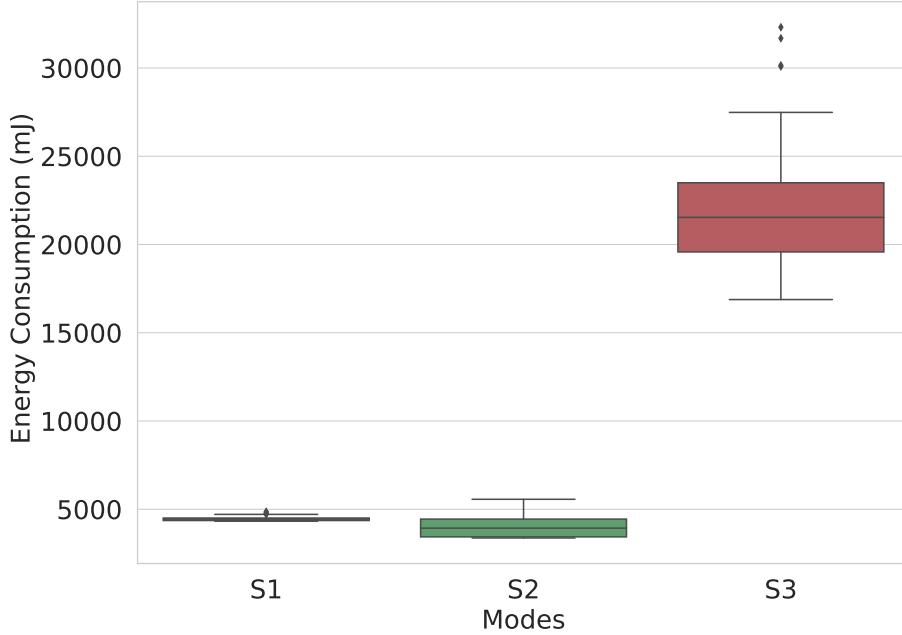


Figure 3.9: Energy variation considering the three cores pinning strategies at 50 % workload

considerable variation-wise, as the variation is 30 times lower in the first strategy ($STD_{S1} = 116mJ, STD_{S3} = 3,452mJ$). This shows that the usage of the hyper-threads technology is not the main reason behind the variation, the first strategy has even less variation than the second one and still uses the hyper-threading.

The reason for the S1 low energy consumption is that one of the two sockets is idle and will likely be in a lower power P-state, even with the disabled C-states. The S2 case is also low energy consuming because by distributing the threads across all the cores, it completes the task faster than in the other cases. Hence, it consumes less energy. The S3 is a high consuming strategy because both sockets are being used, but only half the cores are active. This means that we pay the energy cost for both sockets being operational and for the experiments taking longer to run because of the recurrent context switching.

Our hypothesis regarding the worst results that we observed when using the third strategy is the recurrent context switching, added to the OS scheduling that could reschedule processes from a socket to another, which invalids the cache usage as a process can not take profit of the socket local L3 cache when it moves from a CPU to another (cf. Figure 3.4).

Moreover, the fact that the variation is 4–5 times higher when using the strategy S2 compared to S1 ($STD_{S1} = 116mJ, STD_{S3} = 575mJ$), gives another reason to believe that swapping a process from a CPU to another increases the variation due to CPU micro differ-

Table 3.2: STD (mJ) comparison for 3 pinning strategies

Strategy	S1	S2	S3
Node 1	88	270	1,654
Node 2	79	283	2,096
Node 3	58	287	1,725
Node 4	51	229	1,334

ences, cache misses and cache coherency. While the mean execution time for the strategy S3 is very high ($MeanTime_{S3} = 46s$) compared to the two other strategies ($MeanTime_{S1} = 11s$, $MeanTime_{S2} = 7s$), we see no correlation between the execution time and the energy variation, as the S1 still give less variations than S2 even if it takes 36 % more time to run.

Table 3.2 reports on additional aggregated results for the STD comparison on four other nodes of the cluster Dahu at 50 %, with the benchmarks LU, CG and EP. In fact, the CPU usage strategy S1 is by far the experimentation mode that gave the least variation. The STD is almost 5 times better than the strategy S2, but is up to 10 % more energy consuming ($Mean_{S1} = 4469mJ$, $Mean_{S2} = 4016mJ$). On the other hand, the strategy S3 is the worst, where the energy consumption can be up to 5 times higher than the strategy S2 ($Mean_{S2} = 4016mJ$, $Mean_{S3} = 21645mJ$) and the variation is much worst (30 times compared to the first strategy). These results allow us to have a better understanding of the different processes-to-PU pinning strategies, where isolating the workload on a single CPU is the best strategy. Using the hyper-threads PU on multiple sockets seems to be a bad recommendation, while keeping the hyper-threading enabled on the machine is not problematic, as long as the processes are correctly pinned on the PU. Our experiments show that running one hyper-thread per core is not always the best to do, at the opposite of the claims of [83].

Processes Threshold To answer the second question regarding the evolution of the energy variation at different levels of CPU usage, we varied the used PU's count to track the EV evolution. Figure 3.10 compares the aggregated energy variation when the C-states are on and off using 2, 4 and 8 processes for the benchmarks LU, CG and EP. This figure confirms that disabling the CPU C-states does not decrease the variation for all the workloads, as we can clearly observe, the variation is increasing along with the number of processes. When running only 2 processes, turning off the C-states reduces the STD up to 6 times, but consumes 20 % more energy ($Mean_{C-states-on} = 10,334mJ$, $Mean_{C-states-off} = 12,594mJ$). This variation is 4 times lower when running 4 processes and almost equal to the C-states on mode when running 8 processes. In fact, running more processes implies to use more CPU cores, which

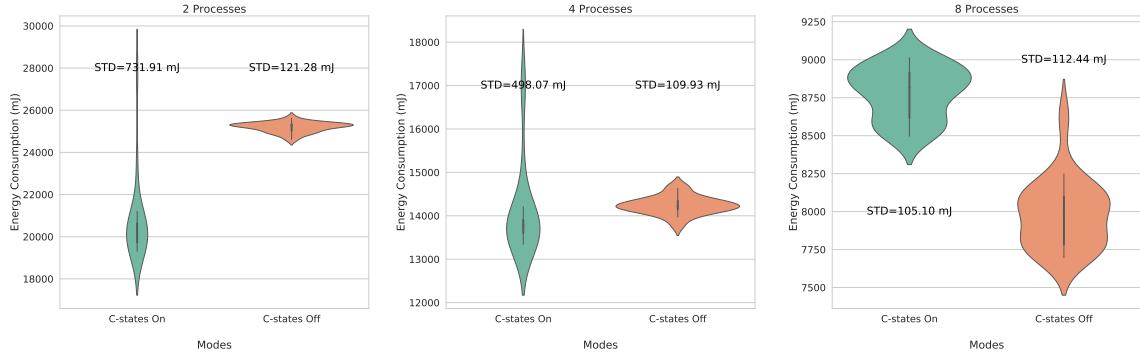


Figure 3.10: C-states effect on the energy variation, regarding the application processes count

reduces the idle cores count, so the cores will more likely stay at a higher consumption state even if the C-states mechanism is on.

In our case, using 4 PU reduces the variation by 4 times and consumes almost the same energy as keeping the C-states mechanism on ($Mean_{C\text{-states}\text{-}on} = 7,048mJ$, $Mean_{C\text{-states}\text{-}off} = 7,119mJ$). This case would be the closest to reality as we do not want to increase the energy consumption while reducing the variation, but using a lower number of PU still results in less variation, even if it increases the overall energy consumption.

We note that disabling the C-states is not recommended in production environments, as it introduces extra energy consumption for low workloads (around 50 % in our case for a single process job). However, our goal is not to optimize the energy consumption, but to minimize the energy variation. Thus, disabling the C-states is very important to stabilize the measurements in some cases when the variation matters the most. Comparing the energy consumptions of two algorithms or two versions of a software systems is an example of use case benefiting from this recommendation.

Turbo Boost The Turbo Boost—also known as *Dynamic Overclocking*—is a feature that has been incorporated in Intel CPU since the Sandy Bridge micro-architecture, and is now widely available on all of the Core i5, Core i7, Core i9 and Xeon series. It automatically raises some of the CPU cores operating frequency for short periods of time, and thus boost performances under specific constraints. When demanding tasks are running, the operating system decides on using the highest performance state of the processor.

Disabling or enabling the Turbo Boost has a direct impact on the CPU frequency behavior, as enabling it allows the CPU to reach higher frequencies in order to execute some tasks for a short period of time. However, its usage does not have a trivial impact on the energy variation. Acun *et al.* [3] tried to track the Turbo Boost impact on the Ivy Bridge and the

Table 3.3: STD (mJ) comparison when enabling/disabling the Turbo Boost

Turbo Boost	Enabled	Disabled
EP / 5 %	310	308
CG / 25 %	95	140
LU / 25 %	204	240
EP / 50 %	84	79
EP / 100 %	125	110

Sandy Bridge architectures. They concluded that it is one of the main responsible for the energy variation, as it increases the variation from 1 % to 16 %. In our study, we included a Turbo Boost experiment in our testbed, to check this property on the recent Xeon Gold processors, covering various workloads.

The experiment we conducted showed that disabling the Turbo Boost does not exhibit any considerable positive or negative effect on the energy variation. Table 3.3 compares the STD when enabling/disabling the Turbo Boost, where the columns are a combination of workload and benchmark. In fact, we only got some minor measurements differences when switching on and off the Turbo Boost, and where in favor or against the usage of the Turbo Boost while repeating tests, considering multiple nodes and benchmarks. This behavior is mainly related to the *thermal design power* (TDP), especially at high workloads executions. When a CPU is used at its maximum capacity, the cores would be heating up very fast and would hit the maximum TDP limit. In this case, the Turbo Boost cannot offer more power to the CPU because of the CPU thermal restrictions. At lower workloads, the tests we conducted proved that the Turbo Boost is not one of the main reasons of the energy variation. In fact, the variation difference is barely noticeable when disabling the Turbo Boost, which cannot be considered as a result regarding the OS activity and the measurement error margin. We cannot affirm that the Turbo Boost does not have an impact on all the CPU, as we only tested on two recent Xeon CPU (clusters Chetemi and Dahu). We confirmed our experiments on these machines 100 times at 5 %, 25 %, 50 % and 100 % workloads.

We conclude that CPU features **highly impact** the energy variation as an answer for RQ 2.

RQ 3: Operating System

The *operating system* (OS) is the layer that exploits the hardware capabilities efficiently. It has been designed to ease the execution of most tasks with multitasking and resource sharing. In some delicate tests and measurements, the OS activity and processes can cause a significant overhead and therefore a potential threat to the validity. The purpose behind this

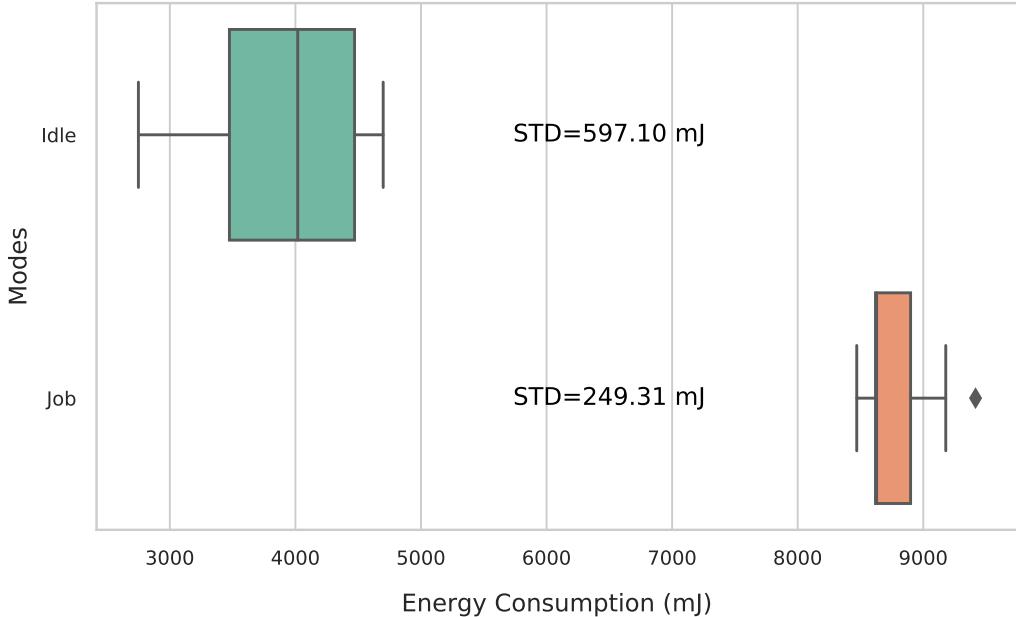


Figure 3.11: OS consumption between idle and when running a single process job

experiment is to determine if the sampled consumption can be reliably related to the tested application, especially for low-workload applications where CPU resources are not heavily used by the application.

The first way to do is to evaluate the OS idle activity consumption, and to compare it to a low workload running job. Therefore, we ran 100 iterations of a single process benchmark EP, LU and CG on multiple nodes from the cluster Dahu, and compared the energy behavior of the node with its idle state on the same duration. The aggregated results, illustrated in Figure 3.11, depict that the idle energy variation is up to 140 % worst than when running a job, even if it consumes 120 % less energy ($Mean_{Job} = 8,746mJ$, $Mean_{Idle} = 3,927mJ$). In fact, for the three nodes, randomly picked from the cluster Dahu, the idle variation is way more important than when a test was running, even if it is a single process test on a 32-cores node. This result shows that OS idle consumption varies widely, due to the lack of activity and the different CPU frequencies states, but it does not mean that this variation is the main responsible for the overall energy variation. The OS behaves differently when a job is running, even if the amount of available cores is more than enough for the OS to keep his idle behavior when running a single process.

Inspecting the OS idle energy variation is not sufficient to relate the energy variation to the active job. In fact, the OS can behave differently regarding the resource usage when running a task. To evaluate the OS and the job energy consumption separately, we used

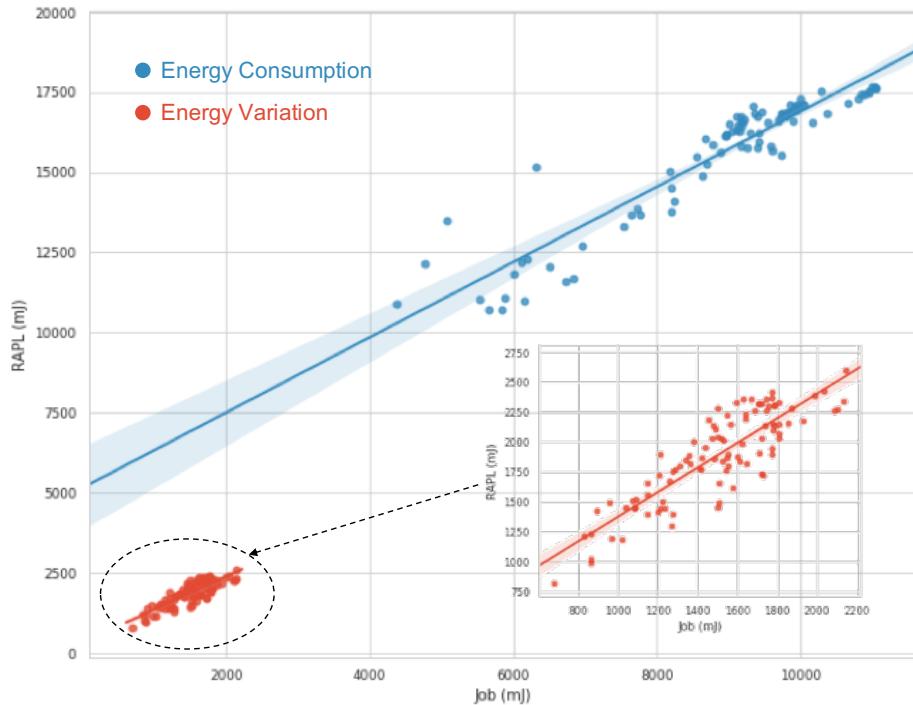


Figure 3.12: The correlation between the RAPL and the job consumption and variation

the POWERAPI toolkit. This fine-grained power meter allows the distribution of the RAPL global energy across all the Cgroups of the OS using a power model. Thus, it is possible to isolate the job energy consumption instead of the global energy consumption delivered by RAPL. To do so, we ran tests with a single process workload on the cluster Dahu, and used the POWERAPI toolkit to measure the energy consumption. Then, we compared the job energy consumption to the global RAPL data. We calculated the Pearson correlation [1] of the energy consumption and variation between global RAPL and POWERAPI, as illustrated in Figure 4.11. The job energy consumption and variation are strongly correlated with the global energy consumption and variation with the coefficients 93.6 % and 85.3 %, respectively. However, this does not completely exclude the OS activity, especially if the jobs have tight interaction with the OS through the signals and system calls. This brings a new question on whether applying extra-tuning on a minimal OS would reduce the variation? As well as what is the effect of the Meltdown security patch—that is known to be causing some performance degradation [69, 79]—on the energy variation?

OS Tuning An OS is a pack of running processes and services that might or not be required its execution. In fact, even using a minimal version of a Debian Linux, we could list many OS running services and process that could be disabled/stopped without impacting

Table 3.4: STD (mJ) comparison before/after tuning the OS

Node	EP	CG	LU
N1	1370 -9 %	78 +7 %	128 +2 %
N2	1278 -7 %	64 -1 %	120 +9 %
N3	1118 +1 %	83 +2 %	93 +7 %

the test execution. This extra-tuning may not be the same depending on the nature of the test or the OS. Thus, we conducted a test with a deeply-tuned OS version. We disabled all the services/processes that are not essential to the OS/test running, including the OS networking interfaces and logging modules, and we only kept the strict minimum required to the experiment’s execution. Table 3.4 reports on the aggregated results for running single process measurements with the benchmarks CG, LU and EP, on three servers of the cluster Dahu, before and after tuning the OS. Every cell contains the *STD* value before the tuning, plus/minus a ratio of the energy variation after the tuning. We notice that the energy variation varies less than 10 % after the extra-tuning. We argue that this variation is not substantial, as it is not stable from a node to another. Moreover, 10 % of variation is not a representative difference, due to many factors that can affect it as the CPU temperature or the measurement errors.

Speculative Executions Meltdown and Spectre are two of the most famous hardware vulnerabilities discovered in 2018, and exploiting them allows a malicious process to access others processes data that is supposed to be private [69, 79]. They both exploit the speculative execution technique where a process anticipates some upcoming tasks, which are not guaranteed to be executed, when extra resources are available, and revert those changes if not. Some OS-level patches had been applied to prevent/reduce the criticality of these vulnerabilities. On the Linux kernel, the patch has been automatically applied since the version 4.14.12. It mitigates the risk by isolating the kernel and the user space and preventing the mapping of most of the kernel memory in the user space. Nikolay *et al.* have studied in [115] the impact of patching the OS on the performance. The results showed that the overall performance decrease is around 2–3 % for most of the benchmarks and real-world applications, only some specific functions can meet a high performance decrease. In our study, we are interested in the applied patch’s impact on the energy variation, as the performance decrease could mean an energy consumption increase. Thus, we ran the same benchmarks LU, CG ad EP on the cluster Dahu with different workloads, using the same OS, with and without the security patch. Table 3.5 reports on the STD values before disabling the security patch. A minus means that the energy varies less without the patch being applied, while a plus means

that it varies more. These results help us to conclude that the security patch's effect on the energy variation is not substantial and can be absorbed through the error margin for the tested benchmarks. In fact, the best case to consider is the benchmark LU where the energy variation is less than 10 % when we disable the security patch, but this difference is still moderate. The little performance difference discussed in [69, 79] may only be responsible of a small variation, which will be absorbed through the measurement tools and external noise error margin in most cases.

Table 3.5: STD (mJ) comparison with/without the security patch

Node	EP	CG	LU
N1	269 +2 %	83 +1 %	108 -6 %
N2	195 +1 %	84 -5 %	121 -9 %
N3	223 +/-1 %	72 -4 %	117 +8 %
N4	276 +3 %	60 +0 %	113 -3 %

To answer RQ 3, we conclude that the OS **should not be the main focus** of the energy variation taming efforts.

RQ 4: Processor Generation

Intel microprocessors have noticeably evolved during these last 20 years. Most of the new CPU come with new enhancements to the chip density, the maximum Frequency or some optimization features like the C-states or the Turbo Boost. This active evolution caused that different generations of CPU can handle a task differently. The aim of this experiment is not to justify the evolution of the variation across CPU versions/generations, but to observe if the user can choose the best node to execute her experiments. Previous papers have discussed the evolution of the energy consumption variation across CPU generations and concluded that the variation is getting higher with the latest CPU generations [Wang et al., 83], which makes measurements stability even worse. In this experiment, we therefore compare four different generations of CPU with the aim to evaluate the energy variation for each CPU and its correlation with the generation. Table 3.6 indicates the characteristics of each of the tested CPU.

Table 3.6 also shows the aggregated energy variation of the different generations of nodes for the benchmarks LU, CG and EP. The results attest that the latest versions of CPU do not necessarily cause more variation. In the experiments we ran, the nodes from the cluster Paranoia tend to cause more variation at high workloads, even if they are from the latest generation. While the Skylake CPU of the cluster Dahu cause often more energy

Table 3.6: STD (mJ) comparison of experiments from 4 clusters

Cluster	Dahu	Chetemi	Ecotype	Paranoia
Arch	Skylake	Broadwell	Broadwell	Ivy Bridge
Freq	3.7 GHz	3.1 GHz	2.9 GHz	3.0 GHz
TDP	125 W	85 W	55 W	95 W
5%	364	210	75	76
50%	98	86	49	244
100%	119	116	106	240

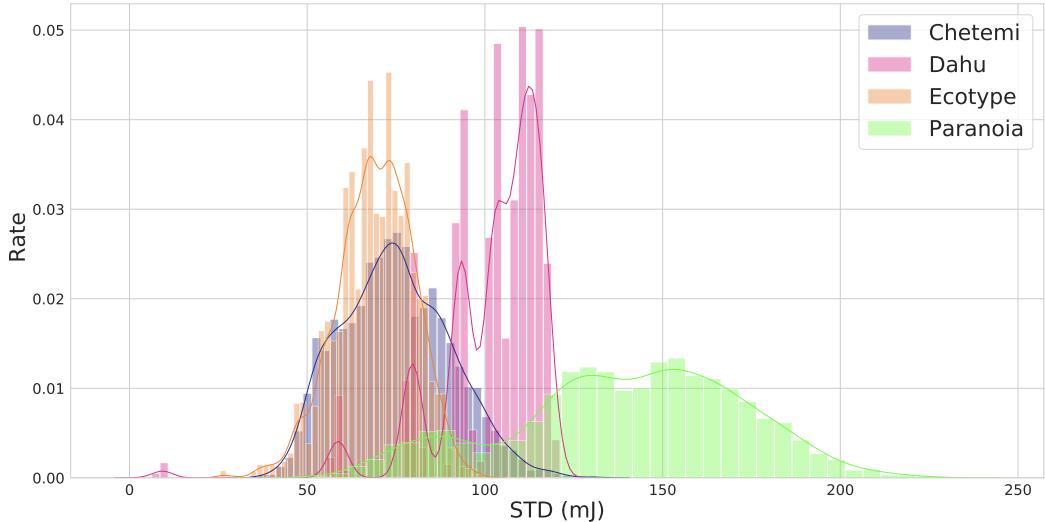


Figure 3.13: Energy consumption STD density of the 4 clusters

variation than Chetemi and the Ecotype Broadwell CPU. We argue that the hypothesis "*the energy consumption on newer CPU varies more*" could be true or not depending on the compared generations, but most importantly, the chips energy behaviors. On the other hand, our experiments showed the lowest energy variation when using the Ecotype CPU, these CPU are not the oldest nor the latest, but are tagged with "L" for their low power/TDP. This result rises another hypothesis when considering CPU choice, which implies selecting the CPU with a low TDP. This hypothesis has been confirmed on all the Ecotype cluster nodes, especially at low and medium workloads.

Figure 3.13 is an illustration of the aggregated STD density of more than 5,000-random values sets taken from all the conducted experiments. This shows that the cluster Paranoia reports the worst variation in most cases, and that Ecotype is the best cluster to consider to get the least variations, as it has a higher density for small variation values.

We conclude on **affirming RQ 4**, as selecting the right CPU can help to get less variations.

3.3.4 Experimental Guidelines

To summarize our experiments, we provide some experimental guidelines in Table 3.7, based on the multiple experiments and analysis we did. These guidelines constitute a set of minimal requirements or best practices, depending on the workload and the criticality of the energy measurement precision. It therefore intends to help practitioners in taming the energy variation on the selected CPU, and conduct the experiments with the least variations.

Table 3.7: Experimental Guidelines for Energy Variations

Guideline	Load	Gain
Use a low TDP CPU	Low & medium	Up to 3×
Disable the CPU C-states	Low	Up to 6×
Use the least of sockets in a case of multiple CPU	Medium	Up to 30×
Avoid the usage of hyper-threading whenever possible	Medium	Up to 5×
Avoid rebooting the machine between tests	High	Up to 1.5×
Do not relate to the machine idle variation to isolate a test EC, the CPU/OS changes its behavior when a test is running and can exhibit less variation than idle	Any	—
Rather focus the optimization efforts on the system under test than the OS	Any	—
Execute all the similar and comparable experiments on a same machine. Identical machines can exhibit many differences regarding their energy behavior	Any	Up to 1.3×

Table 3.7 gives a proper understanding of known factors, like the C-states and its variation reduction at low workloads. However, it also lists some new factors that we identified along the analysis we conducted in Section, such as the results related to the OS or the reboot mode. Some of the guidelines are more useful/efficient for specific workloads, as showed in our experiments. Thus, qualifying the workload before conducting the experiments can help in

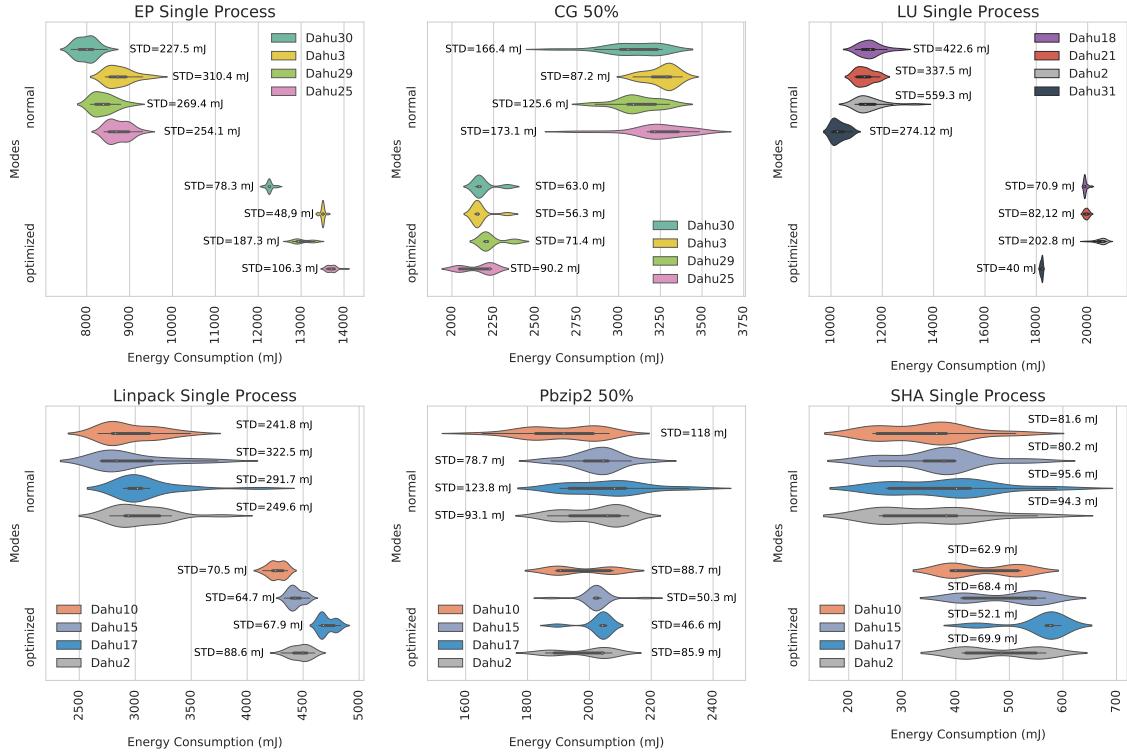


Figure 3.14: Energy variation comparison with/without applying our guidelines

choosing the proper guidelines to apply. Other studied factors are not been mentioned in the guidelines, like the Turbo Boost or the Speculative execution, due to the small effect that has been observed in our study.

In order to validate the accuracy of our guidelines among a varied set of benchmarks on one hand, and their effect on the variation between identical machines on the other hand, we ran seven experiments with benchmarks and real applications on a set of four identical nodes from the cluster Dahu, before (normal mode where everything is left to default and to the charge of the OS) and after (optimized) applying our guidelines. Half of these experiments has been performed at a 50 % workload and the other half on single process jobs. The choice of these two workloads is related to the optimization guidelines that are mainly effective at low and medium workloads. We note that we used the cluster Dahu over Ecotype to highlight the guidelines effect on the nodes where the variation is susceptible to be higher.

Figure 3.14 and 3.15 highlight the improvement brought by the adoption of our guidelines. They demonstrate the intra-node STD reduction at low and medium workloads for all the benchmarks used at different levels. Concretely, for low workloads, the energy variation is

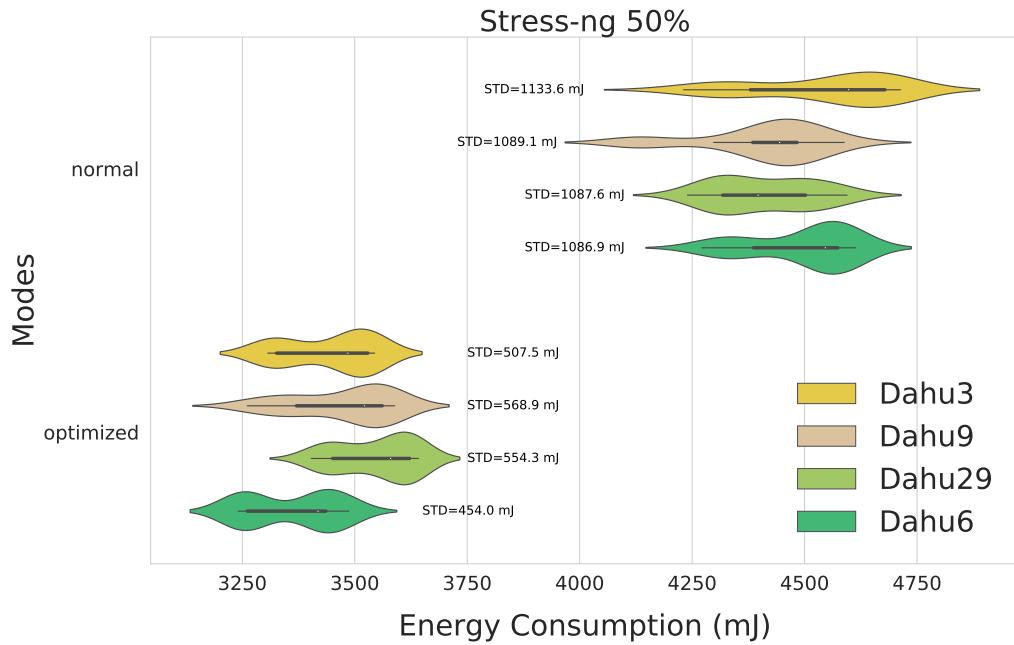


Figure 3.15: Energy variation comparison with/without applying our guidelines for STRESS-NG

2–6 times lower after applying the optimization guidelines for the benchmarks LU and EP, as well as LINPACK, while it is 1.2–1.8 times better for Sha256. For this workload, the overall energy consumption after optimization can be up to 80 % higher due to disabling the C-states to keep all the unused cores at a high power consumption state ($Mean_{LU-normal-Dahu2} = 11,500mJ$, $Mean_{LU-optimized-Dahu2} = 20,508mJ$). For medium workloads, the STD, and thus variation, is up to 100 % better for the benchmark CG, 20–150 % better for the pbzip2 application and up to 100% for STRESS-NG. We note that the optimized version consumes less energy thanks to an appropriate core pinning method.

Figures 3.14 and 3.15 also highlight that applying the guidelines does not reduce the inter-nodes variation in all the cases. This variation can be up to 30 % in modern CPU [Wang et al.]. However, taming the intra-node variation is a good strategy to identify more relevant mediums and medians, and then perform accurate comparisons between the nodes variation. Even though, using the same node is always better, to avoid the extra inter-nodes variation and thus improve the stability of measurements.

3.3.5 Threats to Validity

A number of issues affect the validity of our work. For most of our experiments, we used the Intel RAPL tool, which has evolved along Intel CPU generations to be known as one of the most accurate tools for modern CPU, but still adds an important overhead if we adopt a sampling at high frequency. The other fine-grained tool we used for measurements is POWERAPI. It allows to measure the energy consumption at the granularity of a process or a Cgroup by dividing the RAPL global energy over the running processes using a power model. The usage of POWERAPI adds an error margin because of the power model built over RAPL. The RAPL tool mainly measures the CPU and DRAM energy consumption. However, even running CPU/RAM intensive benchmarks would keep a degree of uncertainty concerning the hard disk and networking energy consumption. In addition, the operating system adds a layer of confusion and uncertainty.

The Intel CPU chip manufacturing process and the materials micro-heterogeneity is one of the biggest issues, as we cannot track or justify some of the energy variation between identical CPU or cores. These CPU/cores might handle frequencies and temperature differently and behave consequently. This hardware heterogeneity also makes reproduction complex and requires the usage of the same nodes on the cluster with the same OS.

3.3.6 Conclusion

In this part, we conducted an empirical study of controllable factors that can increase the energy variations on platforms with some of the latest CPU, and for several workloads. We provide a set of guidelines that can be implemented and tuned (through the OS GRUB for example), especially with the new data centers isolation trend and the cloud usage, even for scientific and R&D purposes. Our guidelines aim at helping the user in reducing the CPU energy variation during systems benchmarking, and conduct more stable experiments when the variation is critical. For example, when comparing the energy consumption of two versions of an algorithm or a software system, where the difference can be tight and need to be measured accurately.

Overall, our results are not intended to nullify the variability of the CPU, as some of this variability is related to the chip manufacturing process and its thermal behavior. The aim of our work is to be able to tame and mitigate this variability along controlled experiments. We studied some previously discussed aspects on some recent CPU, considered new factors that have not been deeply analyzed to the best of our knowledge, and constituted a set of guidelines to achieve the variability mitigating purpose. Some of these factors, like the

C-states usage, can reduce the energy variation up to 500 % at low workloads, while choosing the wrong cores/PU strategy can cause up to 30× more variability.

We believe that our approach can also be used to study/discover other potential variability factors, and extend our results to alternative CPU generations/brands. Most importantly, this should motivate future works on creating a better knowledge on the variability due to CPU manufacturing process and other factors.

3.3.7 Extension

We propose to extend the benchmarking framework suggested by the *Collaboratory on Experimental Evaluation of Software and Systems in Computer Science*⁷ to address those limitations. Instead of only presenting the four primary aspects of their guidelines that were mentioned previously in section 2.1.2, we propose an abstract model to describe an empirical experiment:

context: the hardware and the software configuration for the actual experiment, the purpose of this part is to provide extra information to help readers have a better judgement on the results and to be able reproduce the experiment while diminishing the impact of the external factor . An *orchestrator* which its job to run the experiment. The orchestrator provides three interfaces:

Workload interface: provides a set of functions that should be implemented by the workload and called by the each candidate in the experiment. *Observer interface*: provides a set of metrics that are collected by the orchestrator during the experimentation. note that some of those metrics are external, such the energy consumption of the system , execution time, execution stack ..etc .

Figure 3.16 explains in detail the relationship within each component of the protocol. First of all, we should start with the orchestrator which is the kernel of the experiment. Most of the time this kernel should represent the constants parts of the experiment. The orchestrator is responsible for the following for the coordination between the workload and the candidate, therefore it provides an interface that is implemented by each workload. Later this interface will be provided to each candidate by the orchestrator. Like this we will remove any bias of dependency between the workload and the candidates, which will allow us to extend both parts without any change in the code. Instead of retrieving the data directly from the experiment, we suggest to use an observer to allow the extension of the metrics, this observer in general should be outside the experiment to be able to gather as many informations without impacting the the code.

⁷<http://evaluate.inf.usi.ch/>

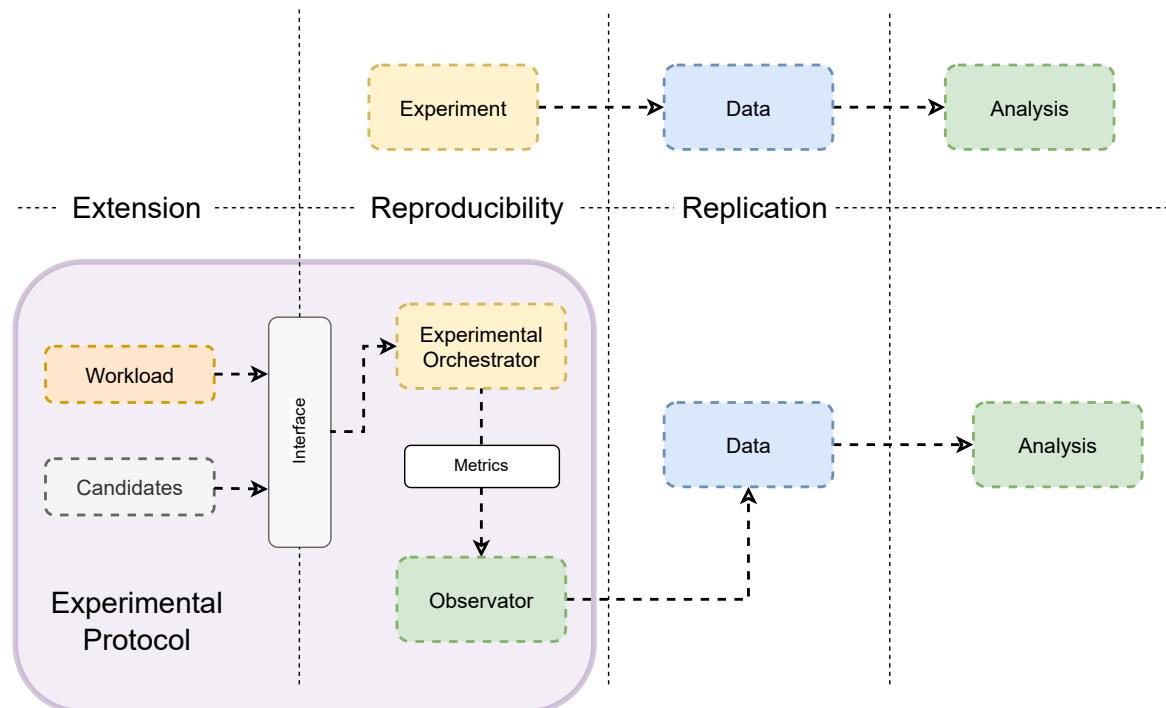


Figure 3.16: Benchmarking protocol

By the end of this study, we have gathered enough guidelines to make the tests more reproducible, accurate and extensible. We created a set of new tests named **energy tests** which are more similar to performance tests. Thanks to the work of two interns [Mamadou and Adrien] we created a CI/CD platform to measure the energy consumption of Java projects and we could track the evolution of the this energy across different stages of the project. In the figure below we see an example of this plugin. For more details please visit the gitlab repository^{8 9}.

⁸<https://gitlab.inria.fr/mamdiall/J-Joules>

⁹<https://gitlab.inria.fr/mamdiall/jjoules-plugin>

Chapter 4

Energy Footprint of Programming Languages

In this chapter, we talk about how the programming language affects how much energy the software consumes. We suggest starting with general micro-benchmarking and watching how each programming language performs with the CPU and memory. The main goal of this chapter is to advise developers on how to choose a programming language based on their project's needs in order to make their product use the least amount of energy possible. For such a question, no answer is obvious. Nonetheless, there are some features we can take from each programming language, such as:

- performance,
- community support,
- scalability,
- energy consumption,
- memory usage.

As we saw in the last chapter, one of the most important things about a test is how well it REPRESENTS the production environment. Therefore, we extend this study to include real-world use cases, with two case studies provided in the parts that follow.

4.1 Investigating Remote Procedure Call Frameworks

4.1.1 Introduction

With the emergence of cloud technologies, many protocols wanted to take the lead. Nowdays, most of the architectures are based on multi-services and microservices. And, to have higher versatility of developers multiple companies choose to be open to different programming languages. Basically, it would be more efficient if we take advantage of each programming language to satisfy a specific need. However, the challenge nowadays is to make the bridge between those platforms. We have many initiatives, such as OpenAPI, that try to create a taxonomy for RESTful APIs. Other approaches implement all the different interfaces of the protocol by themselves, such as RPC.

4.1.2 Research Questions

In this section, we first explore the ease of implementation of this protocol then we will try to answer the following research questions:

RQ 1: How do RPC implementations consume with regards to the size of the incoming request?

RQ 2: How do RPC implementations consume with regards to the number of concurrent clients?

4.1.3 Experimental Protocol

Measurement Context

Hardware settings All the experiments are run on the cluster paravance of the G5K platform. This cluster is composed of 72 identical machines, each one is equipped with 2 Intel Xeon E5-2630 V3, with 128 GB of RAM. For more accuracy, our SUT (*System Under Test*) is equipped with a minimal version of Debian 9 (4.9.0 kernel version), which enforces the core processes required for the purpose of our experiment. Furthermore, we used Docker containers technology for reproducibility of the experiments and the isolation of the servers.

Client and server environments To limit the impact of the network on the experiments, we run both the client and the server on the same machine. However, we isolate each part on a separate socket, in order to reduce the effect that the client might have on the server and *vise versa*. To do so, for each iteration, we always run the same client on socket 0 and the

server that we want to test on socket 1. Both the server and the client use the whole socket for their experiment. In addition, all the additional services, such as the kernel and HwPC sensor, are run on socket 0. Therefore, the only process being executed in socket 1 is the server that we benchmark and monitor.

Key Performance Metrics

Energy measurements To report on the energy consumption, we used HwPC sensor [TOCITE], which is based on Intel RAPL technology, one of the most accurate tools to measure the energy consumption of the CPU and DRAM [TOCITE]. For better accuracy, we ran the HwPC sensor with a frequency of 10 Hz, and we used the same machine for all the experiments in order to reduce the variability [TOCITE].

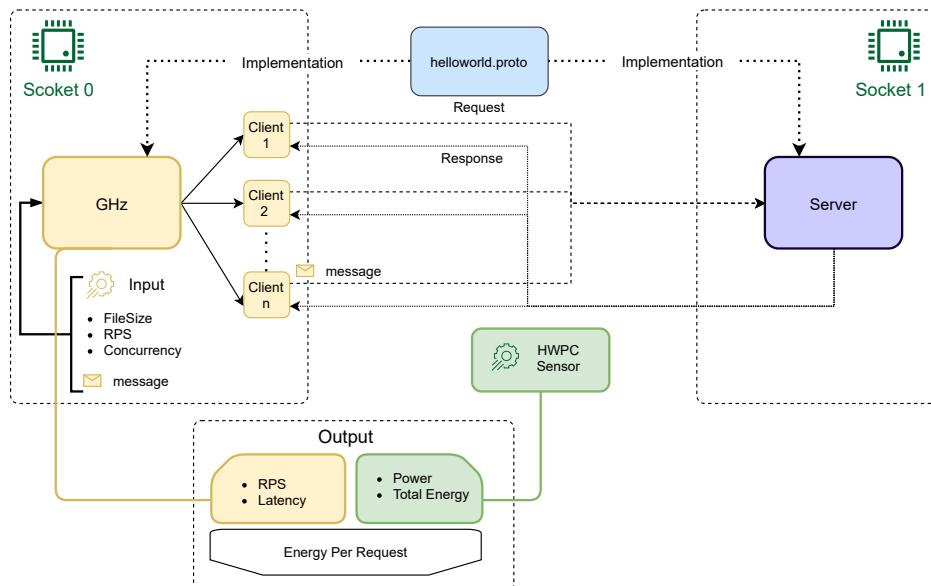


Figure 4.1: Experimental software architecture

performances For better accuracy and more details, We use an updated version of the open source RPC benchmarking tool, named GHZ (<https://ghz.sh/>). The modified version allows us to monitor the average power for each request from both the server and the client sides. The new version is available in the repository¹.

¹https://github.com/chakib-belgaid/energy_ghz

Workload

The purpose of the experiment is to analyse the behaviour of different GRPC implementations. therefore we have two kind of workloads

number of clients : this is the number of concurrent clients that we want to test. It is automatically handled with the ghz client

payload : the size of each request . this varries from 50 Bytes up to 10 MB.

The client consumes the protocol description which can be found if the file `helloworld.proto` to generate an implementation for the message and then forks multiple instances that send the same request to the server

candidates

The server implementations are based on the official implementation by Google for most of the languages. Each server uses 16 cores and is limited to 512 MB of RAM. each implementation is created through a docker image, which will allow us to add new implementations easily.

Extension

: for the sake of extedind the experement we provide a github repository ² that containes the implementation of the experement. Adding **candidates** can be achieved by creating a new docker image and putting it in a new folder named after the language. As for the **workload** it can be extended easly by adding new files in the folder **payload**.

4.1.4 Results & Findings

[RQ 1:] How do RPC implementations react to the size of the data ? The purpose of this question is to study the behaviour of the server when transferring large objects. To do so, we send 80,000 requests to the server whose size scales from 10 bytes up to 10 Megabytes, which results in 10,000 requests per size per server. To eliminate extra factors, we let the server handle the rate at which it can answer each request. However, we put a 20 seconds timeout limit for each request. Therefore, our boundary condition is only the number of requests received by the server. For this experiment, we investigate 4 observable variables :

²https://github.com/chakib-belgaid/energy_benchmarking_grpc2

1. the average power consumption during the process: this will indicate the overall behaviour of the server in working mode for long durations,
2. the tail latency for the 99th percentile, which indicates how performant is the server,
3. the average number of requests per second, which indicates the average number of clients that the server can handle,
4. the average energy cost of a single request: unlike the first indicator, this one shows how green is the implementation taking performance into consideration.

The above figure depicts the overall behaviour of each framework based on the size of request (payload). For each framework, we can distinguish three modes, and they all depend on the payload size:

1. Stress free mode when the server has enough resources to satisfy the requests because they require a memory less than a certain threshold (depends on the language and the platform),
2. Escalation mode when the requests tend to be bigger, however the server can still manage to handle them, and here where we can see a change in the energetic and performance behaviour,
3. Broken state mode when the requests are much heavier and the server break—like 10 MB.

Stress Free Mode

In this mode, the compiled languages tend to consume less resources (average power). JVM-based languages tend to consume more energy, especially Scala. However, we do not observe the same behaviour when it comes to efficiency. Unlike the other interpreted programming languages, PHP performances could be compared to the compiled ones, such as CPP or GO, and even better to some others, such as Swift. JVM-based languages tend to have better performances than the interpreted ones. Furthermore, OpenJDK has shown more efficiency than GraalVM []. Overall, we can have 3 groups when it comes the cost of each request:

- Energy-efficient class: C++, GO, RUST, ELIXIR, and PHP,
- Middle class: Most of the interpreted languages and VM-based ones,
- Energy-greedy class: Crystal and Scala.

Escalation Mode

In this mode, the behaviour of the server depends on the payload. We observe three behaviours:

1. Drop in performances without an increased power, such as .Net core, Java micro-naut, Crystal, and Dart. In this case, the server keeps using the same resources, and sometimes less, because it takes more time to handle the less requests. This class of languages tend to be the most energy consuming when it comes to the cost per request,
2. Increase in power without affecting the performances: such as Go, .Net. The energy consumption of a single request, is affected slightly but still increase,
3. Increase in power and drop in performances: Despite the increase of the power consumption, the server becomes slightly slower, which increases the cost of the energy cost per request. This cost is still better than the first case, which concludes that the servers in the first category are on the verge of breaking.

We can mention the case of Elixir that kept scaling despite the lack of performances compared to other compiled languages (Go, CPP).

Broken state mode

Only four of the 25 configurations could parse the 10 MB files, and only 1 from those could achieve a 76% acceptance rate which is Elixir, the other 3 had less than 3% success rate (Rust, Swift and Dart). The rest could be divided into two categories:

- Timeout where requests took too much time that the client canceled them, in this category we find most of dynamic codes, such as OpenJDK and Kotlin,
- Size of request exceeded the maximum size when the implementation could not handle requests with large size, as observed with .Net, Go, .Net core, CPP, PHP, Scala, Nodejs, Ruby, Python.

[RQ 2:] How do RPC implementations react to the number of clients ?

Power behaviour

Based on the heatmap, we can distinguish two modes:

- Low number of clients when the number of concurrent clients is below 100,
- Moderate to high number of clients when the number of clients exceeds 100.

Lite mode The benchmarked implementations can be grouped into two classes:

1. Energy-efficient frameworks where most of the framework's power consumption is around 33 Watts.
2. Energy-greedy frameworks where the average power consumption is higher than 37 Watts.

In each programming category, we observe both energy-efficient and energy-greedy behaviours. Therefore, we conclude that it depends more on the implementation of the library itself, rather than the category of the programming language. Scala and Kotlin are an excellent example to support this hypothesis, as both of them run on the same virtual machine as Java (OpenJDK 16.1). Yet, their average power is 130% higher than the Java implementation.

Stressed mode Although the same classes remained the same, not all the languages had the same evolution and here we can clearly observe a correlation with the category of the programming language rather than the implementation itself. We can clearly highlight that VM-based languages have a significant increase (double) in the average power consumption after they receive more than 100 concurrent clients. Except PHP, all the interpreted and compiled languages preserved their energetic behaviour. Our hypothesis points to the JIT, since it compiles the code and makes it run faster, hence stressing the CPU. A interesting behaviour has been noticed for the GraalVM: the decrease of energy consumption when increasing the number of the clients. This is related to the drop of the performances, which was probably due to the bottleneck situation where the GraalVM could not handle more than 100 clients simultaneously.

Performance Behaviour

In this section, we study only the number of requests per seconds processed by the server without looking at its energy. We consider three observable variables:

- Satisfaction ratio: how many requests have been satisfied among the total requests,
- Request Per Seconds: The number of the requests that have been answered from the server,
- Tail Latency at 99%: one of the best metrics to evaluate the performances of a server.

Satisfaction ratio Most of the considered frameworks satisfy all the requests, by either reducing the number of requests per second or by increasing the processing time. However, there are some frameworks that have chosen a different approach, such as Dart or Scala, where the choice was to keep a certain limit of latency even if not all the requests are answered. Furthermore, we tend to observe this behaviour among other frameworks, such as Python or Asynchronous NodeJs, when the number of the client exceeds 800.

RPS Most of the servers hit their RPS limit after 5 clients and 100 clients for vm based servers, and after this the number keeps constant, which will decrease the average RPS per client. .Net server is the most performant, followed by Java and Go, while Python and Ruby are the least performant.

Tail Latency The increase in the number of requests per second, does not necessarily mean a lower latency. As one can notice in Table ??, until the 1000 clients, Go provides the least latency beside .Net. GraalVM provides the highest latency, on average. However, Dart tends to become slower when we increase the number of clients, until we pass the 600 simultaneous clients, and there it changes its behaviour, instead of satisfying most the requests it notify the clients directly that the server is saturated, hence a drop in satisfaction ratio, and an amelioration for the average latency.

Energy Per Request

Now, after we made a separation between the energy and the performances, we have seen that most of the performance servers tend to be energy-greedy, so we propose to investigate this trade-off between energy and performances. To do so, we report on an average cost of a single request, in Joules. Except GraalVM when the cost of the a single request increases with when we add more clients, all the frameworks report on a constant cost, Java, .net and go are the most energy efficient, while Python and Ruby may cost up to 10x more. Therefore, we conclude that the number of clients does not impact the energy significantly. Then, we study how the payload size of the requests impact the energy consumption of the framework.

4.1.5 Threads to Validity

4.1.6 Conclusion

4.2 Investigating Web Application Frameworks

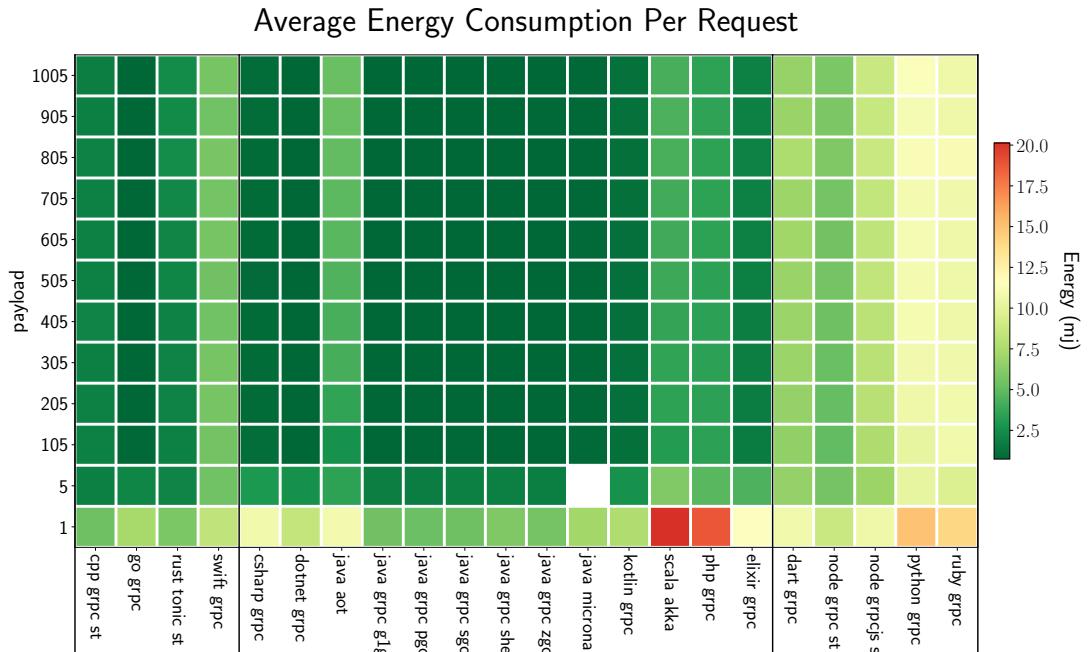


Figure 4.2: Experimental software architecture

4.2.1 Introduction

Nowadays, web applications are dominating online systems. From Google, to Facebook and others, web applications are widely deployed across organizations and continuously accessed by end-users, both for their personal and professional daily tasks. In practice, the development of these web applications heavily rely on a wide ecosystem of *web frameworks*, which are intended to ease and foster the development process. However, once deployed, the applications developed with such web frameworks do not exhibit the same performances, as reported by the *Web Framework Benchmarks* periodically published by the TECHEMPOWER company.³ Thanks to such benchmarks, developers can take informed decisions on the most performing technology to adopt to implement their web applications. Unfortunately, one can regret that developers and benchmark providers focus mostly on popularity and performance criteria when picking a web framework, with less considerations for the resource consumption implications of their choice. This is all the more regrettable that cloud providers are more and more adopted by developers to host these web applications. While cloud providers offer a convenient elastic provision of resources to scale according to application requirements, this convenience may induce critical cost for their business.

³<https://www.techempower.com/benchmarks>

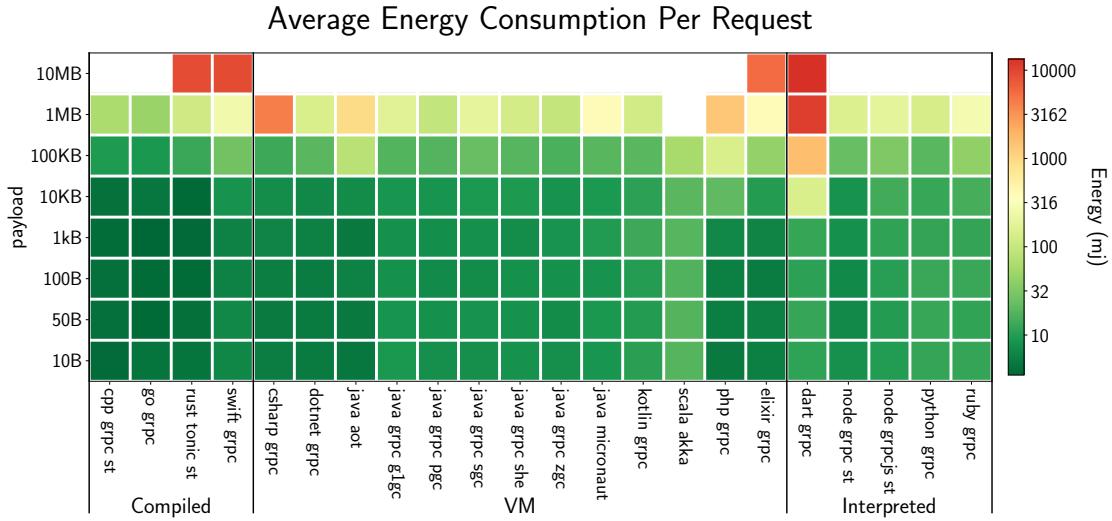


Figure 4.3: Experimental software architecture

Beyond the economical cost of web applications, one can also question the more global impact of web applications on worldwide carbon emissions. Given the tremendous success of web applications, their deployment has severely increased over last years, thus causing a rebound effect on the power consumption of server infrastructure—being hosted or supported by cloud providers. While one can challenge the relevance of features that are continuously deployed by developers to keep engaging end-users, reconciling economical and environmental concerns remains an open challenge to address.

Given this context, this paper intends to contribute to this challenge by investigating the energy footprint of web frameworks. In particular, we aim to support the developers of web applications with relevant guidelines that can help them to choose the web framework that is not only the most popular or provide the best performances, but also exhibits a low energy footprint. By minimizing the energy consumed to process user requests, with no service quality penalty, developers can reduce the operational cost of their web applications and contribute to reduce worldwide carbon emissions of ICT.

To achieve this objective, we leverage the TECHEMPOWER *Web Framework Benchmarks* to incorporate server-side energy measurements obtained from a software-defined power meter, named POWERAPI [1]. These measurements are then analyzed in depth to understand the key criteria that can amper the power consumption of web frameworks and derive

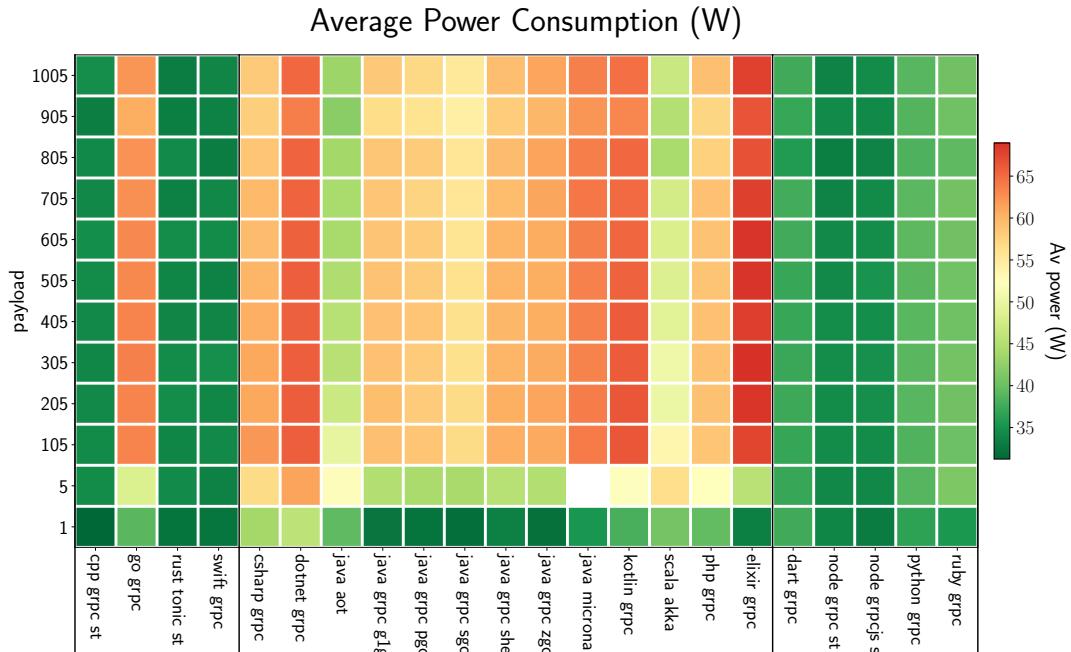


Figure 4.4: Experimental software architecture

guidelines to can help developers to pick the most energy efficient web frameworks according to their requirements.

The remainder of this paper is organized as follows.

4.2.2 Comparaison of Web Frameworks

Many studies have been conducted to compare the performance of web frameworks.

We can cite [46] where they made a comparison between two of the most famous java frameworks, Play and Spring, or the work of [13] when they compare different PHP frameworks using 6 criterions intrinsic durability, industrialized solution, technical adaptability, strategy, technical architecture, and Speed, In the previous paper we find that all those 6 criterions have their own wait when it comes to choose which framework to take for a project. In our study we want to push a 7th criterion that impacts the economic outcome of the project.

4.2.3 Energy Efficiency in Software Engineering

In their paper [107] the authors studied the impact of programming languages on energy, time, and Memory by using the CLBG benchmark, where they executed 10 different benchmarks [

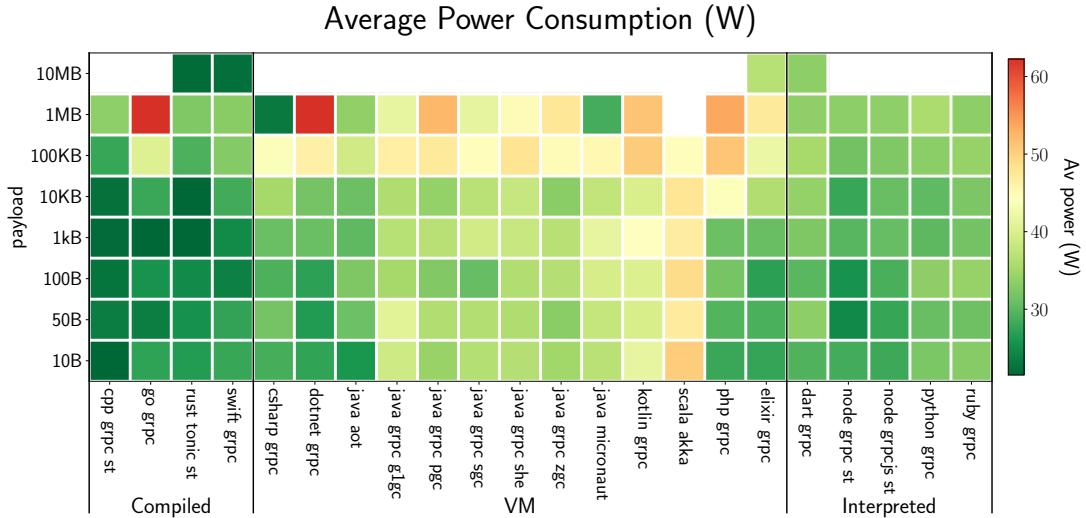


Figure 4.5: Experimental software architecture

add citation] accross 27 wel-know programming languages [add reference] the work of the authors was an extension of the a research marled by [add sitation[6 in the authers paper]]. we continued their work with measuring the impact of programming languages choice in real life application instead of microbenchmark

Irene et al. [82] investigated the impact of web servers on energy when handleling web applications . they analysed 7 applications executed within 4 serveces with 38 different scenarios . the authors showed that the energy greatly depends on the web server , howverver the impact of the application may influence this energitical behaviour of the server.

In their approach they used measured the energy consumption during the integration tests and for us we measure with a simulation of a website when there is a client in another machine that requests the website, therefore we isolate the energy consumption of the server from the the client's one.

Other works have been done on the client side, as an example [108] as they concluded that theres is a variation among the different websites and the impact of the browser on this energy.

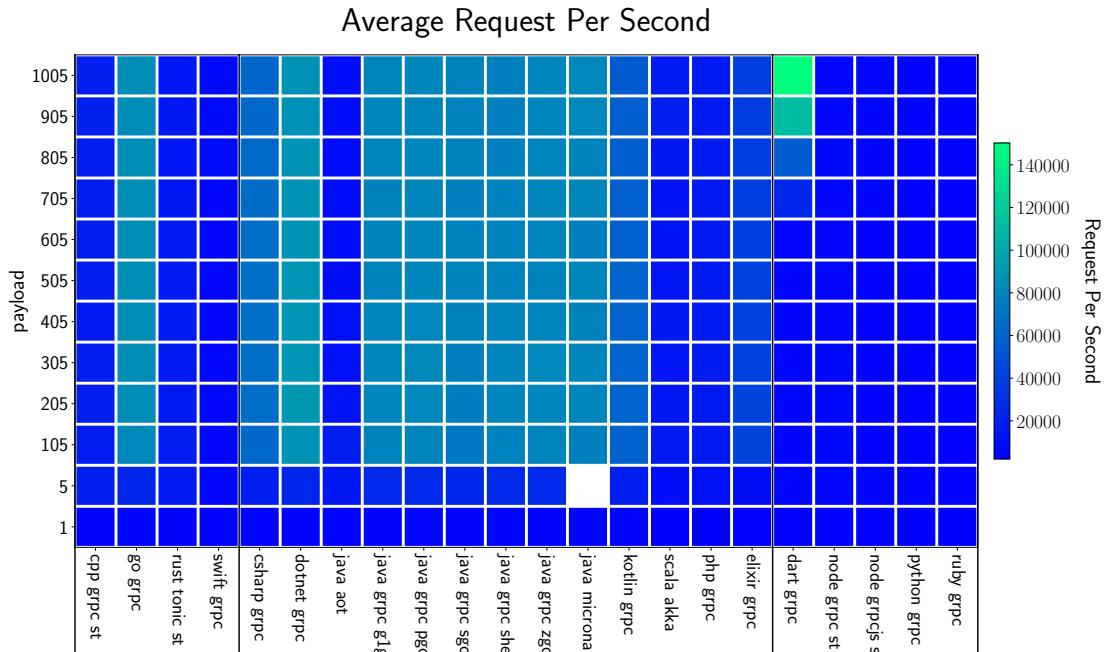


Figure 4.6: Experimental software architecture

4.3 Experimental Protocol

In this section, we describe the environment used during the experiments, covering hardware material, experiments of the framework and the methodology.

4.3.1 Measurement Context

The purpose of this exmeperiment is to highlight the impact of the technology stack of developping a website on the energy consumption while its execution.

Table ?? highlights the number of frameworks used in the experiment and they passed each category of tests. As we see in the table below, some of the frameworks worked on certain conditons while they failed on other tests such as nickel from rust, while nickel might be one of the greenest rust framework, it doens't work with databases. Therefore, it can be used within all the situations, but if the website doesnt interact with a database then it might be the best choice.

many reasons are behind the failure, either there were no implementation or just there were some errors while handling the request.

TODO : we decided to skip the idle part in the validation test since it is not relevant

Table 4.1: the number of framework passed per test

Language	Bb	Query	Update	Plaintext	Fortune	Json	Total
c	1	1	1	6	1	5	15
c#	21	20	14	12	14	17	98
c++	27	16	14	20	13	25	115
cfml	2	1	1	1	1	2	8
clojure	8	8	5	6	7	8	42
common lisp	2	/	/	/	/	2	4
crystal	3	1	/	2	/	2	8
d	3	2	1	2	1	3	12
dart	/	/	/	2	/	2	4
elixir	1	1	/	/	/	1	3
erlang	3	2	/	3	1	3	12
f#	/	/	/	4	2	8	14
go	19	18	16	15	15	19	102
groovy	1	/	/	1	/	2	4
haskell	1	1	1	2	1	2	8
java	20	20	18	26	21	26	131
javascript	19	19	16	14	17	14	99
julia	/	/	/	1	/	1	2
kotlin	10	9	6	5	5	10	45
lua	1	1	/	1	1	2	6
nim	/	/	/	2	/	3	5
ocaml	4	4	3	1	2	5	19
perl	2	/	/	1	/	2	5
php	22	18	15	10	12	14	91
prolog	/	/	/	1	/	1	2
python	31	21	15	17	16	30	130
racket	1	/	/	/	/	/	1
ruby	23	15	11	8	12	19	88
rust	8	7	6	9	8	10	48
scala	7	6	3	8	5	11	40
swift	2	2	/	2	/	2	8
typescript	4	2	2	3	2	6	19
v	/	/	/	1	/	1	2
vala	/	/	/	1	/	2	3
vb	2	2	2	1	2	1	10
total	248	197	150	188	159	261	1203

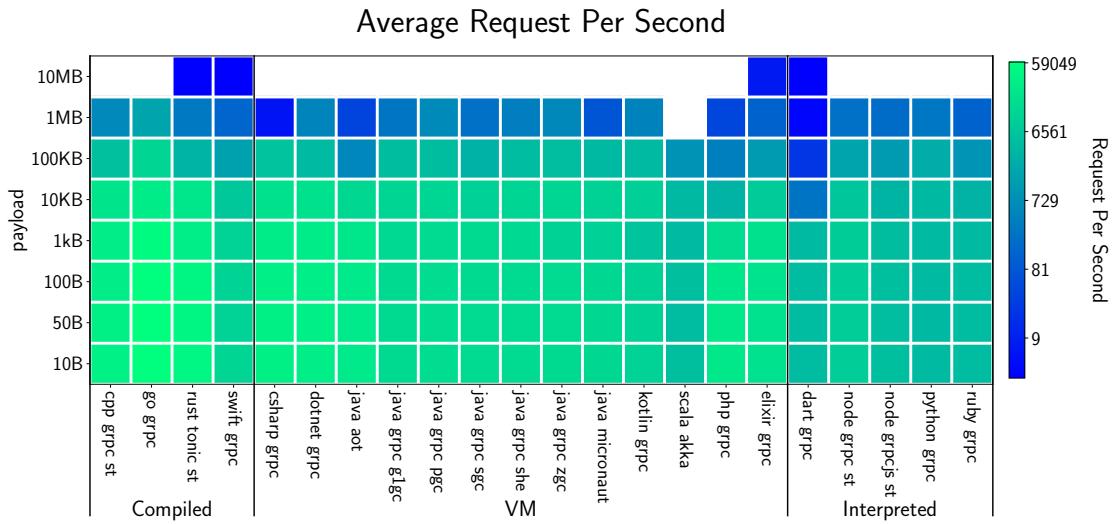


Figure 4.7: Experimental software architecture

- orchestrator : the part that is responsible for creating docker images, selecting the benchmarks and launching the tests;
- web server : or the system under test (SUT) is the machine responsible for launching the framework by means of the pre-installed powermeter;
- database server : offers the same data base that will be used by all the frameworks during the tests
- client machines : To avoid the bottleneck on the client's side, client requests are sent from another machine (one or many) that simulates hundreds of concurrent connections to the framework.
- recorder : the party responsible for collecting the power data from the SUT and the performance metrics collected by the clients ***TODO: add link to the measurement process***

The tests have been executed in machines from the cluster chetemi of the ***ADD Reference: Grid5000*** plateform . ***TODO: add hardware description***

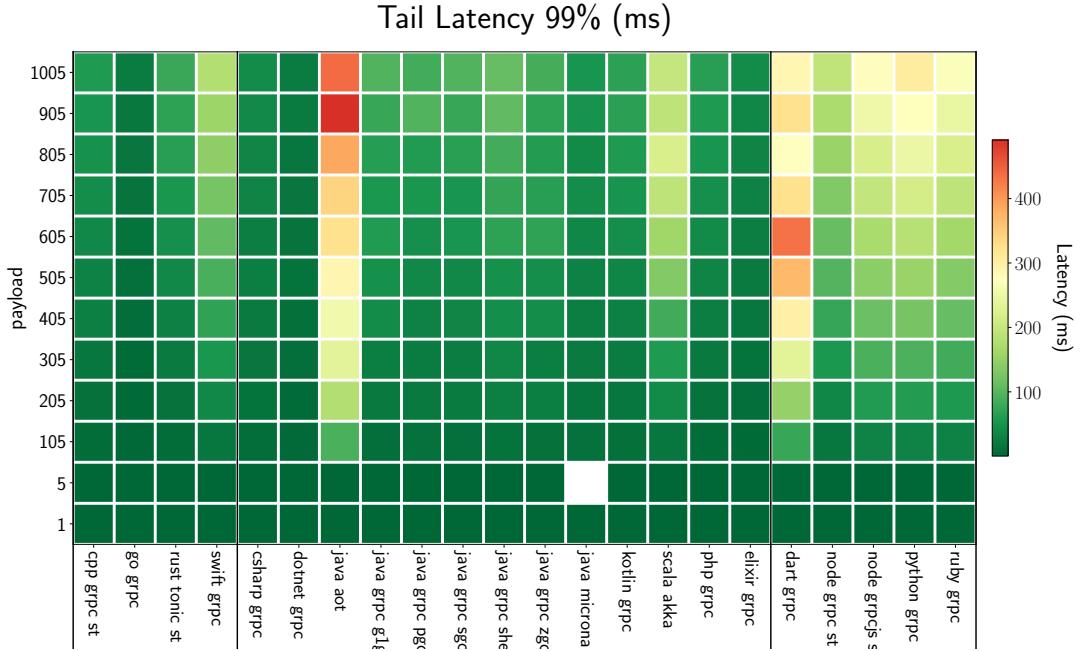


Figure 4.8: Experimental software architecture

Note It has been proven in the work of Eddie Antonio Santos et al. that docker does not influence on the energy consumption[40]. Thus, using its containers and isolation allow to avoid any alteration of the operation system after testing one benchmark and to take advantage of its reproducibility.

4.3.2 Workload

To measure compare the energy consumption and performance efficiency between multiple frameworks, each framework used to implement the same website answering the same URLs and requesting the same database. Then we run the same algorithm for all implementations.

1. launch the framework
2. wait for 20s for the warmup
3. measure the average power when the framework is in idle state
4. using multiple clients. we send the same request in simultaneously during 20s
5. increase the number of parallel request

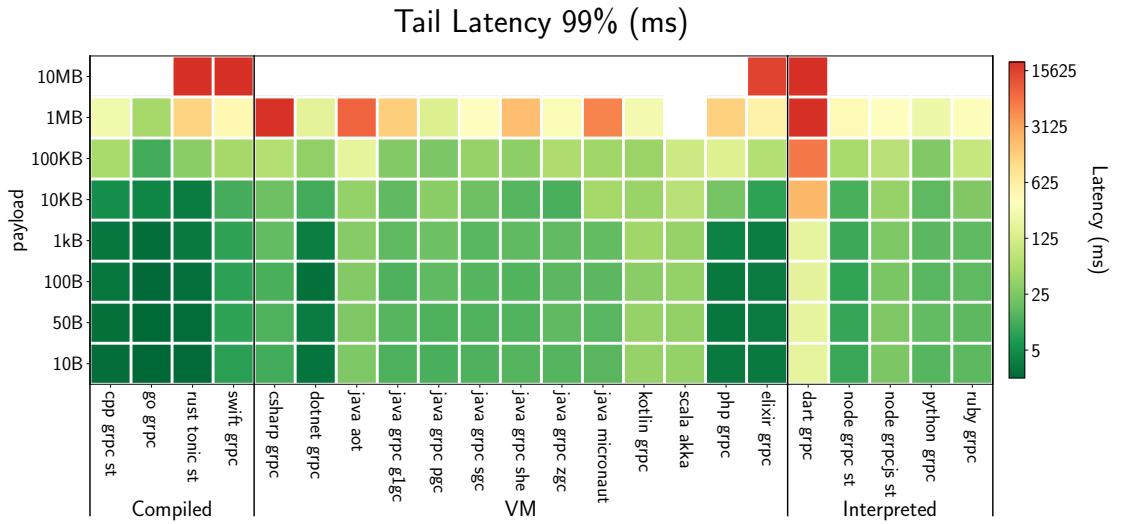


Figure 4.9: Experimental software architecture

6. measure the enrgy during this execution
7. change the request type
8. repeat from 3 rd step

the section will detail each type of experiments and the purpose behind it giving some examples of the expecteed response

Test Categories

We have 7 categories of tests :

Idle In this test we measure the idle energy consumption of the framework: this reflects the average energy consumption of an application during the periods when we have no clients. For example, a company website beyond working hours or a small website at night.

Single Query During this test, each request is processed by fetching a single row from a simple database table. That row is then serialized as a JSON response.

Multiple Queries this test aims to test the behavior of a framework when it processes multiple entries from the database. Therefore each request is processed by fetching multiple rows from a simple database table and serializing these rows as a JSON response. In this case we use the maximum number of clients which is 512512

Fortunes In this test, the framework's ORM is used to fetch all rows from a database table containing an unknown number of Unix fortune cookie messages (the table has 12 rows, but the code cannot have foreknowledge of the table's size). An additional fortune cookie message is inserted into the list at runtime and then the list is sorted by the message text. Finally, the list is delivered to the client using a server-side HTML template. The message text must be considered untrusted and properly escaped and the UTF-8 fortune messages must be rendered properly.

Update queries This test exercises database writes. Each request is processed by fetching multiple rows from a simple database table, converting the rows to in-memory objects, modifying one attribute of each object in memory, updating each associated row in the database individually, and then serializing the list of objects as a JSON response. same as the multiple queries we use here the maximum number of clients which is 512512

The response is analogous to the multiple-query test.

Plain text In this test, the framework responds with the simplest of responses: a "Hello, World" message rendered as plain text. The size of the response is kept small so that gigabit Ethernet is not the limiting factor for all implementations. HTTP pipelining is enabled and higher client-side concurrency levels are used for this test.

JSON Serialization In this test, each response is a JSON serialization of a freshly-instantiated object that maps the key message to the value "Hello, World!".

For each one of the above scenarios, we consider different levels of stress table shows the different levels for each scenario.

Table 4.2: Stress levels for each scenario

Scenario	type of stress	level 1	level 2	level 3	level 4	level 5	level 6	level 7
Single query	Number of parallel clients	16	32	64	128	256	512	/
Multiple queries	Number of rows to read from the database	1	5	10	15	20	30	50
Update queries	Number of rows to update in the database	1	5	10	15	20	30	50
Fortunes	Number of parallel clients	16	32	64	128	256	512	/
JSON Serialization	Number of parallel clients	16	32	64	128	256	512	/
Plain text	Number of parallel clients	256	1024	4096	16384	32384	/	/

To add extra scenarios one might implement a python class that handles the metadata of the workload, such as the query route, the query parameters and the expected results.

4.3.3 Key Performance Metrics

We focus on comparing the energy behavior of different frameworks in multiple scenarios

To measure the energy consumption of those frameworks, we launch each one for a fixed duration, then all the clients send multiple requests simultaneously. We calculate the number of satisfied responses which reflects the performance of the framework, the average latency and the global energy consumed during the whole period to deduce the energy cost of each request.

Runtime Measurements

- energy measurement : we use **ADD Reference: powerapi**, a software powermeter to gather the running power of the SUT, after we project the timestamps of each experiment phase to calculate the energy consumption of the framework during this phase. Energy is an integral of power over time, so we use a numerical approach to estimate this energy. **TODO: add formula** after this we devide the calculated energy per number of responses during this phase we have 4 metrics :
 - Total cost of the energy during each period
 - Total number of requests
 - Average latency
 - Average energy cost per request

Architecture

The aim is to compare the energy consumption of different frameworks. For this we consider the framework as a blackbox and we take into account the response to the 6 previous requests using the same database. In order to isolate the energy consumption of the framework the test is run in a separate machine with the minimum services and the powermeter. Figure 4.10 illustrates the architecture of our system

Bias Analysis we are aware of bias analysis regarding the estimation of the total energy cost , the interference of other system processes during the execution and some external events. Thus, we run experiments multiple times and compute the average.

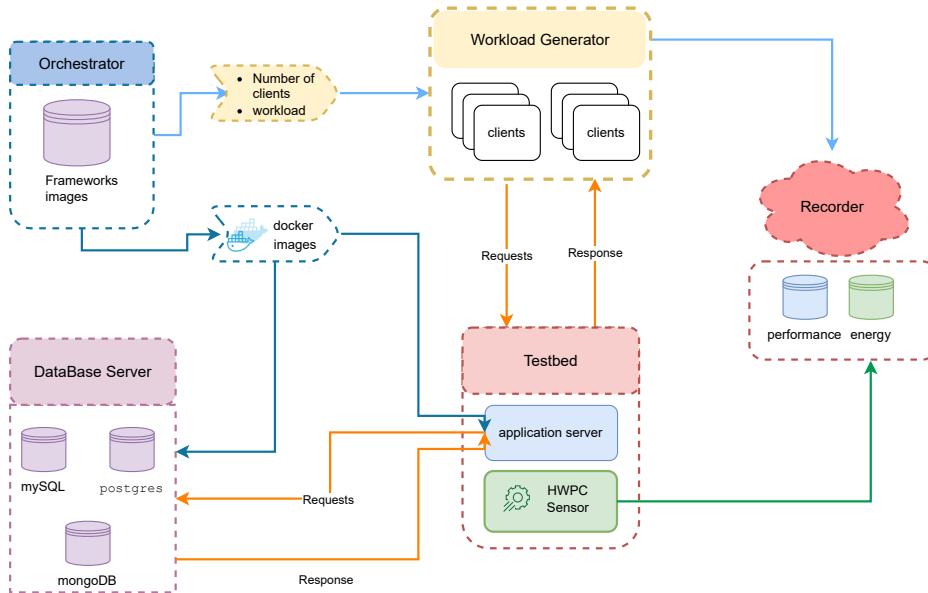


Figure 4.10: Architecture of the experiments

4.3.4 Extension

To follow the guidelines that we presented in chapter 3 we provide a github repository⁴ where one can add extra **candidates** by creating a new project using the option –new. then they have just to fill the template and provide the docker image file.

to configure the **workload** we provide the option –concurrency-levels and –duration. As for the choice of the database it is added in the docker image.

4.4 Results & Findings

Overall we had 8,750 tests, and all can be found in the repository.⁵ In this section, we will use the previous results to answer the following questions

- is there a dominant programming language when it comes to performance, energy consumption, and latency ?
- which class of programming language is performing well ?
- is there a correlation between energy consumption and latency ?
- is there an impact on the server when it comes to changing the database?

⁴<https://github.com/chakib-belgaid/FrameworkBenchmarks>

⁵<https://github.com/chakib-belgaid/frameworks-benchmarks-results>

Since most of the companies use the same stack, we won't be aggregating the results since it may lead to confusion. For example, comparing the average energy consumption of each programming language will not reflect reality, particularly when we have two frameworks on opposite ends of the spectrum.

4.4.1 Overall all statistics

In order to determine which framework/stack is performing well we need to establish some general idea about the average energy consumption and latency of the frameworks. Instead of giving just the raw energy consumption of those frameworks, we will provide some green factors to determine which one is eco-friendly and which one is greedy. In this part, we will discuss the average behavior of the frameworks, highlight some tendencies, and eliminate the outliers. As we said in the threats to validity, being an outlier in this case does not mean that the framework is not performing well; it just means that the framework is not performing well in the same way as the others within the context of this experiment.

On the other hand the cost a single request is proportionate to the number of requests per second of the framework

To narrow the research space, we will look for a correlation between the metrics. The Pearson correlation coefficient [128] will be used to determine this correlation. Because the Shapiro-Wilk [114] test yielded a p-value of 0.0 for all metrics.

Figure 4.11 depicts the correlation between the metrics for the 6 scenarios. The Pearson correlation coefficient quantifies the linear correlation between two variables X and Y. It ranges from -1 to +1, with 1 being total positive linear correlation, 0 representing no linear correlation, and -1 representing total negative linear correlation. The stronger the correlation, the closer the value is to 1 or -1. The weaker the correlation, the closer the value is to zero. The Pearson product-moment correlation coefficient is another name for the correlation coefficient. This correlation coefficient is also known as the Pearson product-moment correlation coefficient. One can notice that there is a strong correlation between the energy consumption of the CPU and the DRAM for most of the scenarios. Moreover the Average energy consumption of DRAM is one sixth of the CPU energy consumption. Therefore, in this study we will focus more on the CPU energy consumption. For more insights about the DRAM consumption we refer the reader to our github repository⁶.

⁶<https://github.com/chakib-belgaid/frameworks-benchmarks-results>

Another strong correlation is between the number of requests per second and the average cost of a single request. This is because the cost a single request is proportionate to the number of requests per second of the framework, since the Average Power consumption remains constant after a certain threshold of clients.

Unlike the Multiple queries and update queries, other scenario depicts a weak correlation between latency and number of requests per second. The reason behind a such anomaly is the fact that we summarized the data when we had a multiple number of clients. If we calculate the correlation when we have a fixed number of client, like the case of update queries (512 clients). one can notice a strong correlation. Figure ?? demonstrates a such behavior. As one can notice. for each level we have a linear clusturing.

Another interesting result is the relationship between

interestingly there is no correlation between the energy consumption of the CPU and the latency, which means that we can achieve both a low latency and a low energy consumption at the same time.

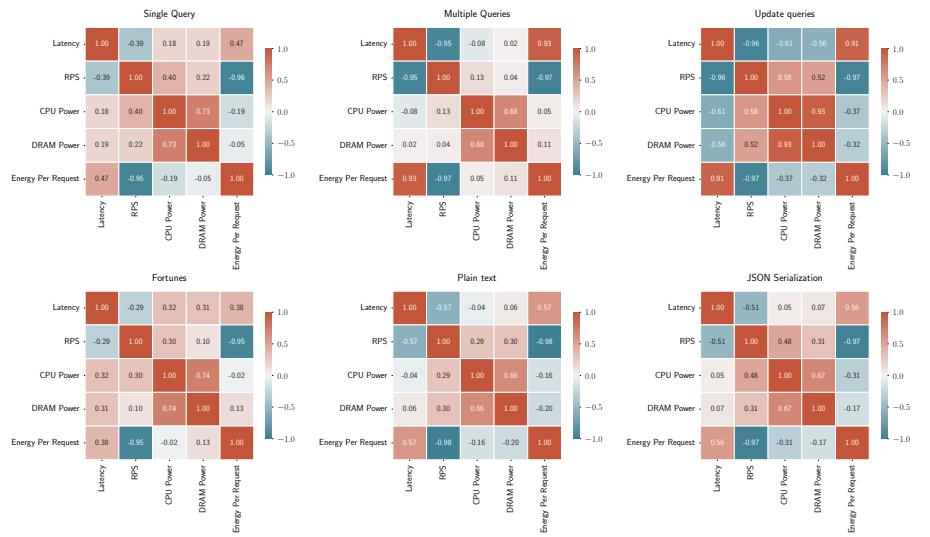


Figure 4.11: Spearman Rank Correlation between different metrics

then we group the implementation by the compiling type ,therefore we will have.

- compiled languages
- interpreted languages
- JVM-based languages
- .net-based languages
- other VM-based languages

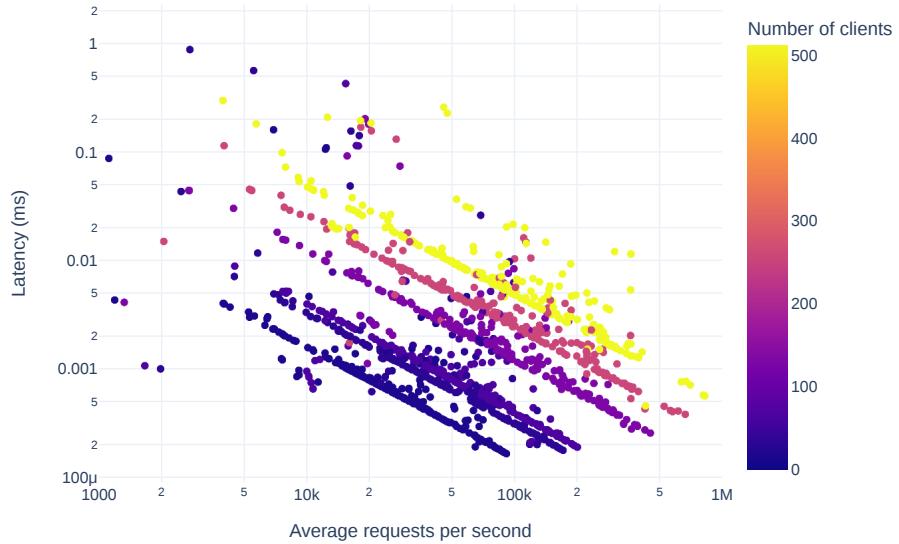


Figure 4.12: correlation of latency and number of requests per second for a single query

4.5 Threats to Validity

We are aware of the bias induced by the implementation of a candidates, therefore we propose the framework (see the part of extension) to allow the readers to confirm themselves any new hypothesis. regarding a new framework , an new workload a new database ..etc

Figure 4.13: Average power consumption for the Single query test

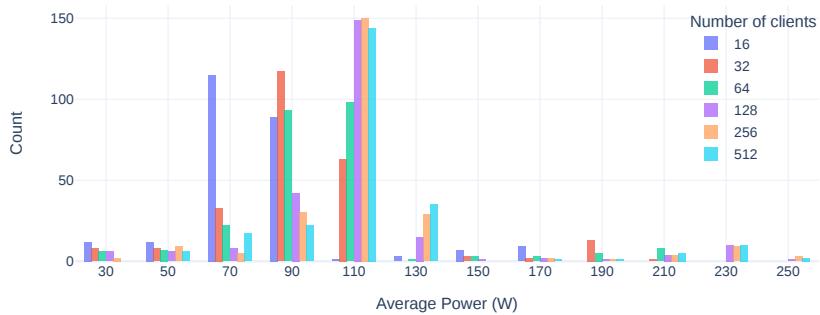


Figure 4.14: total request vs average power consumption for the Single query test (size of circles represents the number of clients)



Figure 4.15: total request vs average power consumption for the multiple queries test (Size of circles represents the size of the query)



Figure 4.16: total request vs average power consumption for the update queries test (Size of circles represents the size of the query)

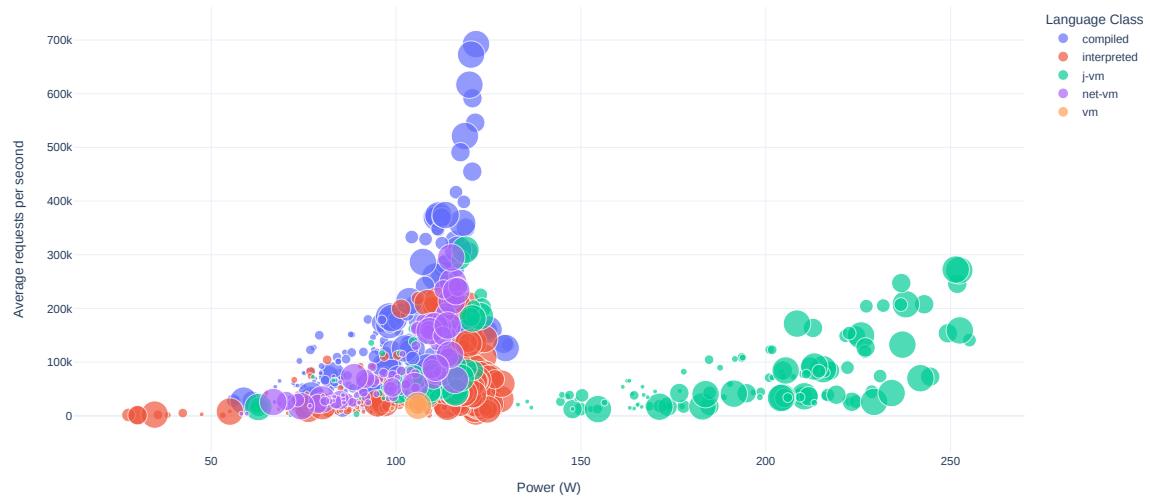


Figure 4.17: total request vs average power consumption for Fortunes test (Size of circles represents the number of clients)

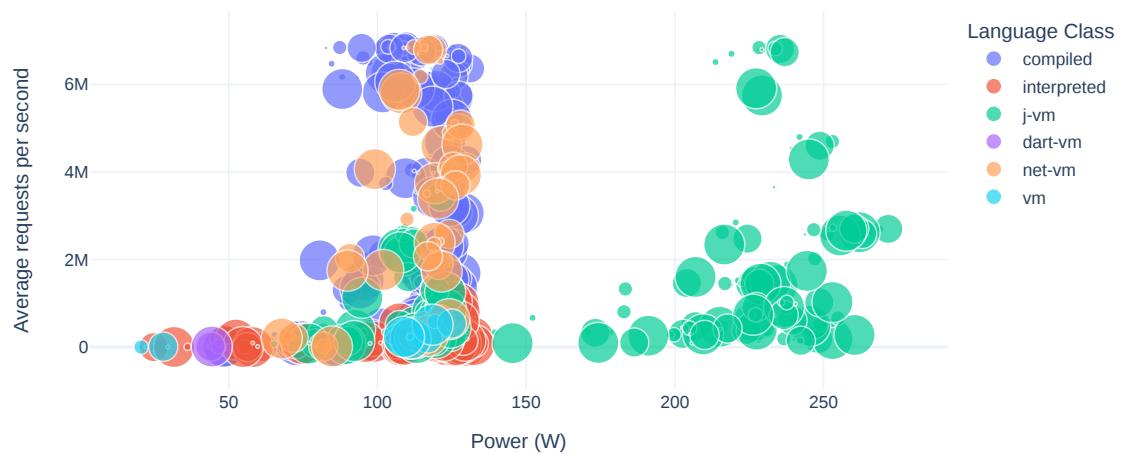


Figure 4.18: total request vs average power consumption for plainText test (size of circles represents the number of clients)



Figure 4.19: total request vs average power consumption for JSON Serialization test (size of circles represents the number of clients)

Part II

Optimizing Application Runtimes

Chapter 5

The Impact of Python Runtime on Energy Consumption

5.1 Introduction

Dynamic programming languages, except Perl, have surpassed compiled programming languages in terms of popularity among software system developers over the past decade (cf. Figure 5.1). However, it remains unclear whether this category of dynamic programming languages can truly compete with compiled ones in terms of power consumption.

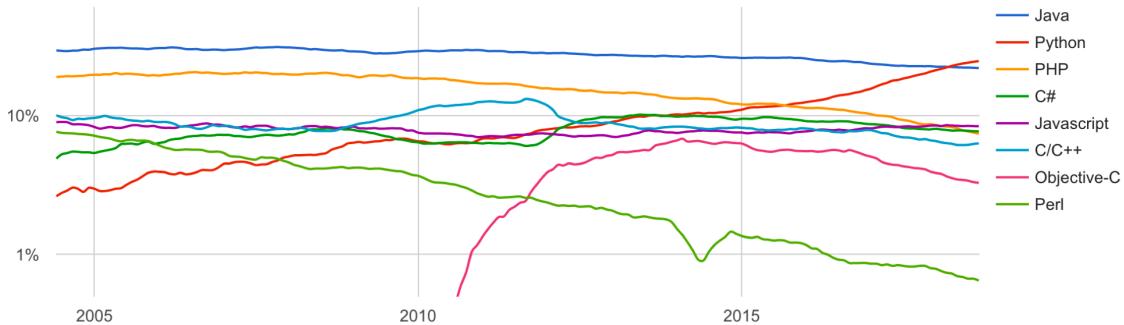


Figure 5.1: PYPL Popularity of Programming Languages [?].

In particular, Noureddine *et al.* [95] in 2012, and then Pereira *et al.* [107] in 2017, conducted empirical power measurements on this topic, and both concluded that compiled programming languages overcome dynamic ones when it comes to power consumption. According to their experiments, an interpreted programming language, like Python, can impose up to a 7,588 % energy overhead compared to C [107] (cf. Figure 5.2). In this chapter, we explore the oblivious optimizations that can be applied to Python legacy applications to

reduce their energy footprint. As Python is widely adopted by software services deployed in public and private cloud infrastructures, we believe that our contributions will benefit a wide diversity of legacy systems and not only favorably contribute to reducing the carbon emissions of ICT, but also reduce their cloud invoice for the resources consumed by these services. More specifically, this chapter focuses on runtime optimizations that can be adopted by developers to leverage the power consumption of Python applications. We start by studying the impact of programmer choices, such as the type of data or control structures, on the global energy consumption of the execution code. Then, we discuss other factors, such as the levels of concurrency, before we investigate other non-intrusive approaches to optimize the energy consumption of applications. One type of optimization includes alternative interpreters and libraries that are dedicated to optimizing the code without changing its structures, such as *ahead-of-time* (AOT) compilation and *just-in-time* (JIT) libraries that are maintained by the community.

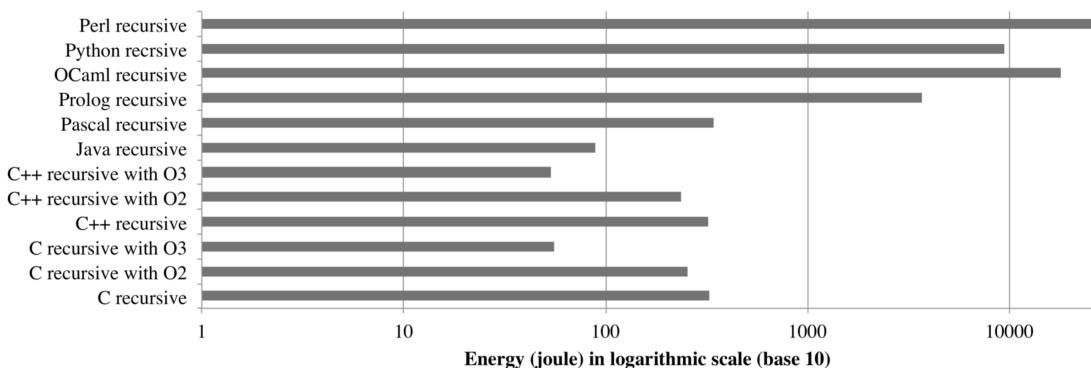


Figure 5.2: Energy consumption of a recursive implementation of Tower of Hanoi program in different languages [95]

5.2 Motivation

5.2.1 Python Popularity

Nowadays, Python seems to attract a large community of developers who are interested in data analysis, web development, system administration, and machine learning. According to a survey conducted in 2018 by JetBrains,¹ one can fear that the wide adoption of dynamic programming languages, like Python, in production may critically hamper the power consumption of ICT. As the popularity of such dynamic programming languages partly builds

¹<https://www.jetbrains.com/research/python-developers-survey-2018/>

on the wealth and the diversity of their ecosystem (*e.g.*, the NumPY, SciKit Learn, and Panda libraries in Python), one cannot reasonably expect that developers will likely move to an alternative programming language mostly for energy considerations. Rather, we believe that a better option consists of leveraging the strength of this rich ecosystem to promote energy-efficient solutions to improve the power consumption of legacy software systems.

5.2.2 Python Gluttony

According to [109] and [95], Python tends to be one more energy hungry programming language. As one can notice in Figure 5.2, Python consumes 30 times more than C or C++. The benchmark was done with an implementation of the Tower of Hanoi² of 30 disks.

Python consumes a lot of energy, mainly because it is slow in execution. Its flexibility and simplicity caused it to drop off in performance because Python gains its flexibility from being a dynamic language. Therefore, it needs an interpreter to execute its programs, which makes them much slower compared to the others that are written in compiled programming languages, such as C and C++ or semi-compiled languages like Java.

As shown in Table 5.1, one can observe that, for most of the applications taken for the *Computer Language Benchmark Game* (CLBG), Python takes more time to execute—the only case that he was not the worst one was in the benchmark `regex-redux` where he beats Go—and in some cases the gap was huge, such as in `n-body` where Python took around 100 times more than C++.³

Table 5.1: Comparison of CLBG execution times (in seconds) depending on programming languages.

	C	C++	Java	Python	Go
pidigits	1.75	1.89	3.13	3.51	2.04
reverse-complement	1.75	2.95	3.31	16.76	4.00
regex-redux	1.45	1.66	10.5	15.56	28.69
k-nucleotide	5.07	3.66	8.66	79.79	15.36
binary-trees	2.55	2.63	8.28	92.72	28.90
fasta	1.32	1.33	2.32	62.88	2.07
Fannkuch-redux	8.72	10.62	17.9	547.23	17.82
n-body	9.17	8.24	22.0	882.00	21.00
spectral-norm	1.99	1.98	4.27	193.86	3.95
Mandelbrot	1.64	1.51	6.96	279.68	5.47

²https://en.wikipedia.org/wiki/Tower_of_Hanoi

³<https://benchmarksgame-team.pages.debian.net/benchmarksgame/index.html>

5.2.3 Python Limits

To reduce the energy consumption of Python, we started by targeting the main usage of this programming language, which is revealed to be data science and web development. Figure 5.3 illustrates a study published by the JetBrains company on Python developers⁴. 57% of the respondents reported that they use Python for data science, 51% said they are using it for web development, and around 40% are using it for system administration. ***TODO : The options in this survey were not mutually exclusive. As a result, the total of the percentages is greater than 100%.***

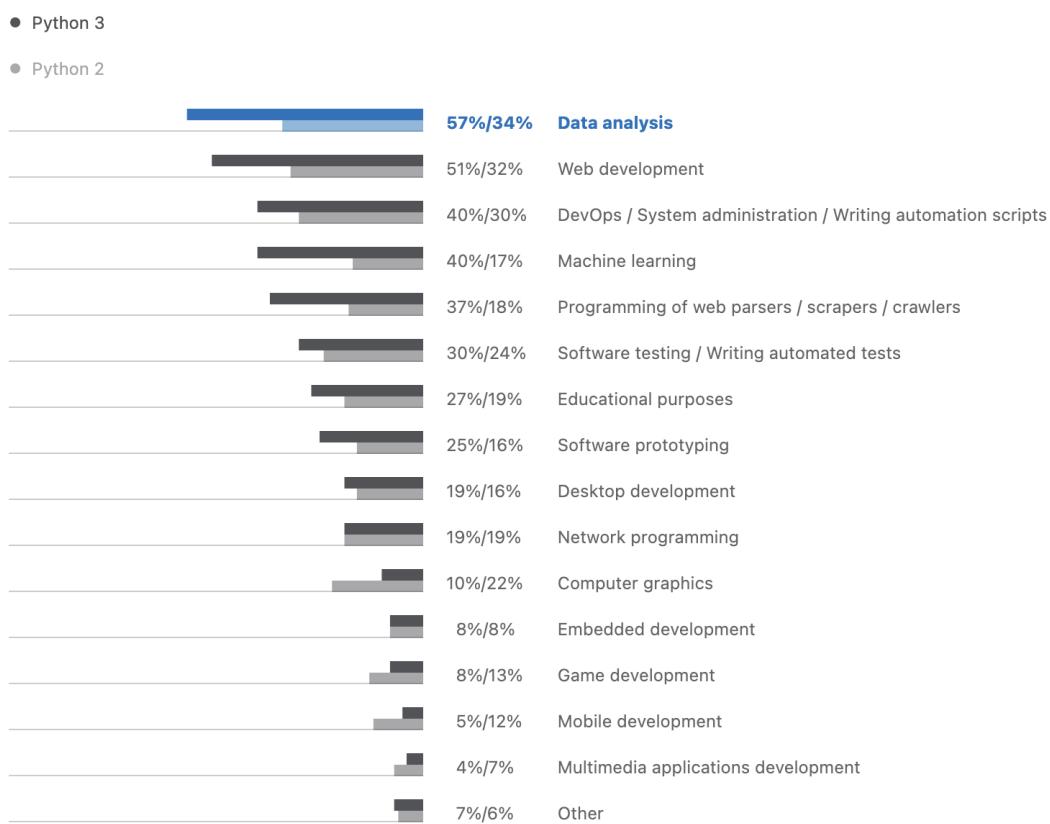


Figure 5.3: Use cases of Python (source: JetBrains).

In this chapter, we investigate the energy footprint of Python in its most popular domains of adoption. We first explore the data and control structures, aiming to reveal some fundamental guidelines, as [?] did in [?]. Then, we measure the energy consumption of several Python implementations to propose a non-intrusive technique to improve energy efficiency. Therefore, this chapter will address the following research questions:

⁴<https://www.jetbrains.com/lp/python-developers-survey-2020>

RQ 1: *What is the energy footprint of Python when used in data science?*

RQ 2: *Are the Python guidelines energy efficient by construction?*

RQ 3: *Can we reduce the energy consumption of Python programs without altering the source code?*

To answer these questions, we report on 4 case studies that intend to answer these research questions. First, we study the energy behavior of Python in two application contexts: web applications (cf. Section 5.3) and machine learning. Then, we dive deeper into the energy consumption of Python core structures, before concluding with the impact of the Python interpreter on energy consumption.

5.3 Green Web Development

Django⁵ and Flask⁶ are the most popular frameworks for web development. According to a Jetbrains poll⁷, Django is used in 40% of the cases, while Flask is used in 41% of the cases.

In contrast to Flask, which is still a microframework, Django is a high-level web framework that provides a standard method for creating and maintaining complex and scalable database-driven websites quickly and effectively. Therefore, this study will focus on The latter one.

5.3.1 life cycle of a request in Django

Django is a MVT (Model-View-Template) framework, which means that it follows the MVC (Model-View-Controller) pattern. Figure 5.4 describes the life cycle of a request in this framework. Whenever a request arrives in Django, it is processed by several middlewares, one at a time. These middlewares are in charge of authentication, security, and so on. Once the request is processed by these middlewares, it is passed to the URL Router, which will simply extract the URL from the request and will try to match it to the defined URLs. After getting the matching URL, the corresponding view function is called. This function is responsible for treating the request, gathering the data, and then generating the response that will be put inside a template to be returned as a response. However, as mentioned before, Django is an MVT model. which offers an automatic way to retrieve data from the database to the view using the *Object Relational Mapping* (ORM) [100].

First, we start by investigating the energy consumption of the request-response life cycle in Django, in ordrer to determine which part consumes the most energy. To do so, we created

⁵<https://www.djangoproject.com/>

⁶<https://flask.palletsprojects.com/>

⁷<https://lp.jetbrains.com/python-developers-survey-2021/>

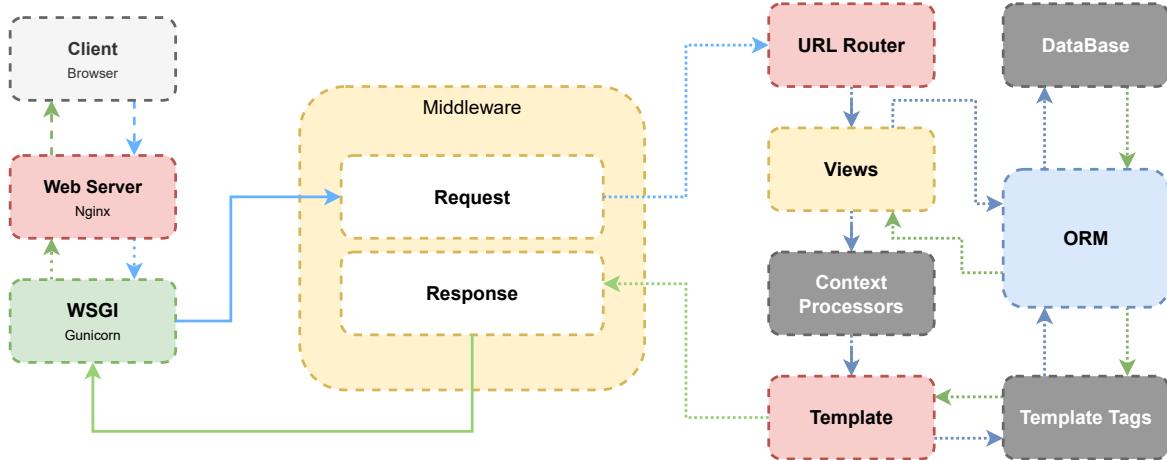


Figure 5.4: Request-Response life cycle in Django

a simple Django application that simply returns the response of the request. and then tracked its energy consumption using joulehunter⁸.

Joulehunter is an open source library that we developed in order to help practitioners find the energy hotspots in their applications using statistical profilers. In the case of Django, it can be added as middleware without any extra setup or changes to the source code. The energy consumption of the request-response life cycle in Django is shown in the figure 5.5. As we can notice, 91.4% of the total energy consumption is spent to resolve the request by retrieving the data while only 5.27% is spent to render the response.

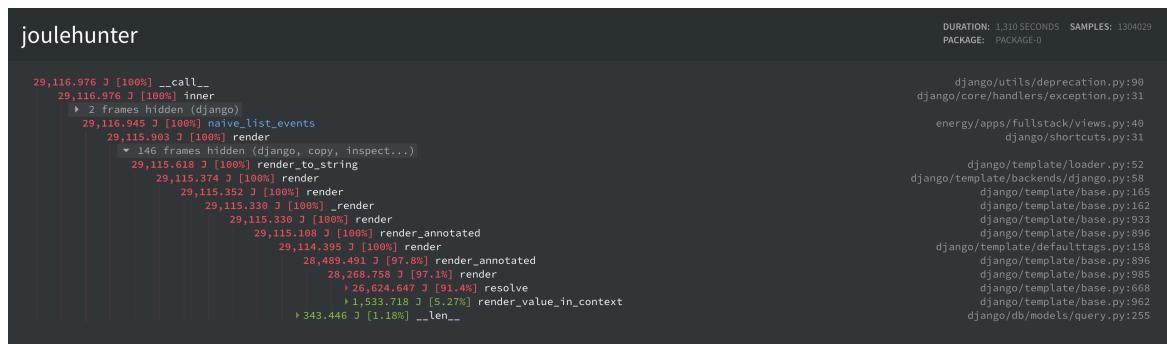


Figure 5.5: Tree representation of the energy consumption of a single request in django (naive version)

Therefore, we chose to observe if the choice of the database and the ORM impacts the energy consumption and the performance of the website. To accomplish so, we examined the cost of a single request of the prior website using various ways to extract data from the

⁸<https://github.com/powerapi-ng/joulehunter>

database. We considered using two different databases, POSTGRESQL and SQLITE3, that store the same records, and three different ways to fetch the records.

1. Vanilla relies only on the ORM to retrieve the data,
2. Prefetch queries the data before being requested,
3. Optimized leverages SQL without passing by the ORM.

As one can observe in Figure 5.6, the strategy to query stored data has a huge impact on the energy consumption. As the Vanilla strategy can consume up to $10\times$ more energy than the Optimized one. Conversely, the choice of the database does not exhibit a key impact on the total energy despite their different behavior regarding the execution time and the average power. This can be useful to support developers in choosing which database engine they can adopt, based on the number of expected requests and the targeted performance.

Another interesting observation is the impact of the interpreter, as Figure 5.6 highlights. For example, using the Pypy interpreter reduces the energy consumption, even when we adopting the Vanilla strategy.

Finally, we run the same experiment with the Optimized strategy using joulehunter. Figure 5.7 shows the new graph of the energy consumption. As one can notice, while the rendering method consumed the same amount of energy as in the Vanilla strategy (arround 1.3KJ), the resolve part droped by $20\times$.

Conclusion We conclude this study by stating that the database and ORM selection have a major impact on the website's performance and energy consumption. And unlike the rendering portion, there are many options available for the database, the ORM, and the data handling strategy.

Figure 5.11 reports on the energy consumption of Python applications when introducing parallelism. Python applications are single-threaded systems constrained by the *global lock system* (GLS). However, due to the increase of cores/threads per CPU, many Python libraries started to take advantage of this hardware feature by allowing concurrent execution of Python processes. For example, the multiprocessing library⁹ spawns sub-processes to increase the degree of concurrency of Python applications. As one can see in Figure 5.11, the energy consumption is correlated with the number of threads until reaching the limit of physical cores, when concurrent processes start to compete for the CPU.

Before reaching the limit of physical cores, we also observe that the scheduler of the operating system tends to favor the execution of processes on the same physical core, by

⁹<https://docs.python.org/3/library/multiprocessing.html>

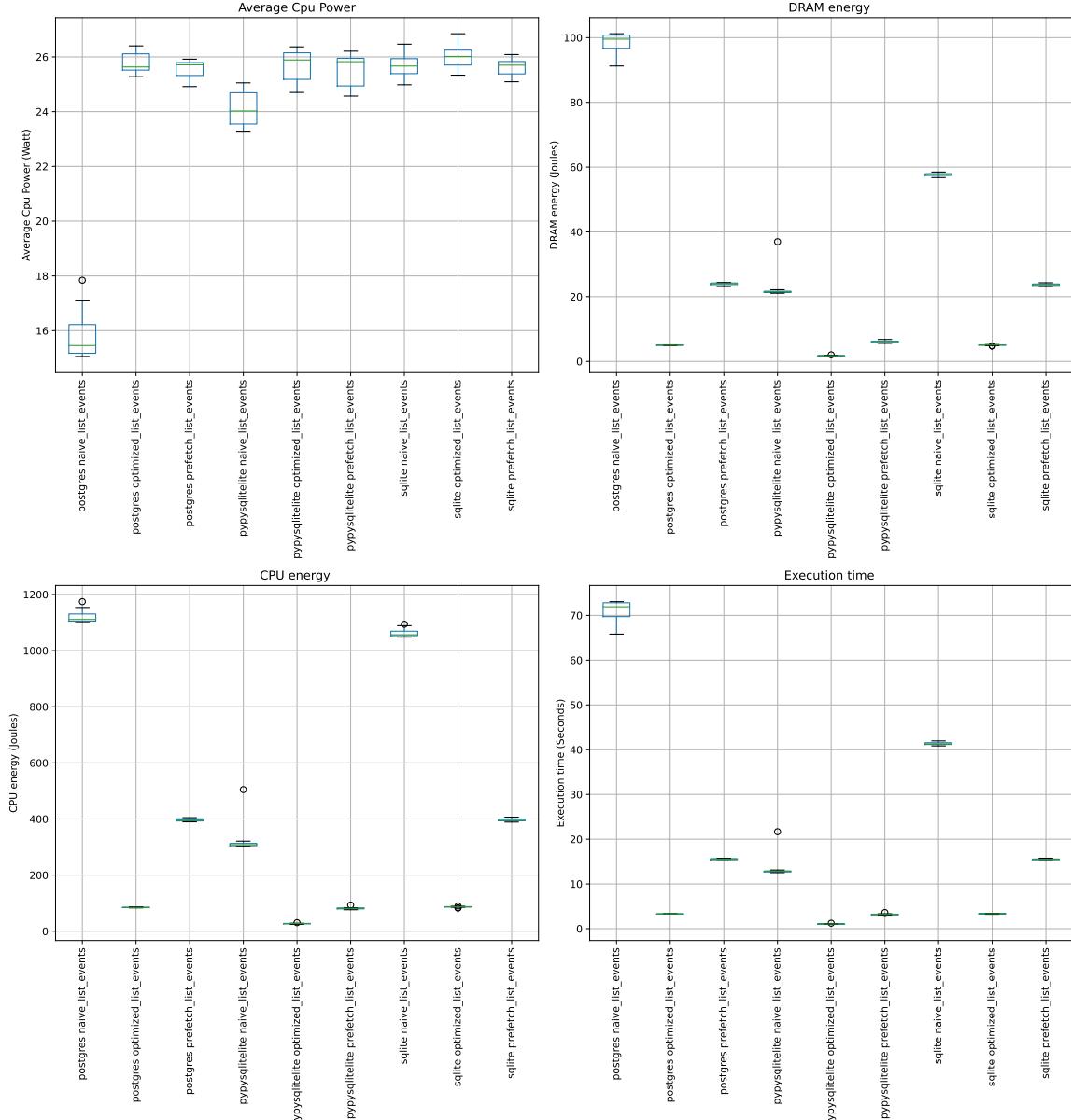


Figure 5.6: Energy behavior resulting from data access strategies.

taking advantage of the hyper-thread feature. While this strategy aims to save energy by leveraging the ACPI P-states and C-states of unallocated cores, this leads to increase execution times.

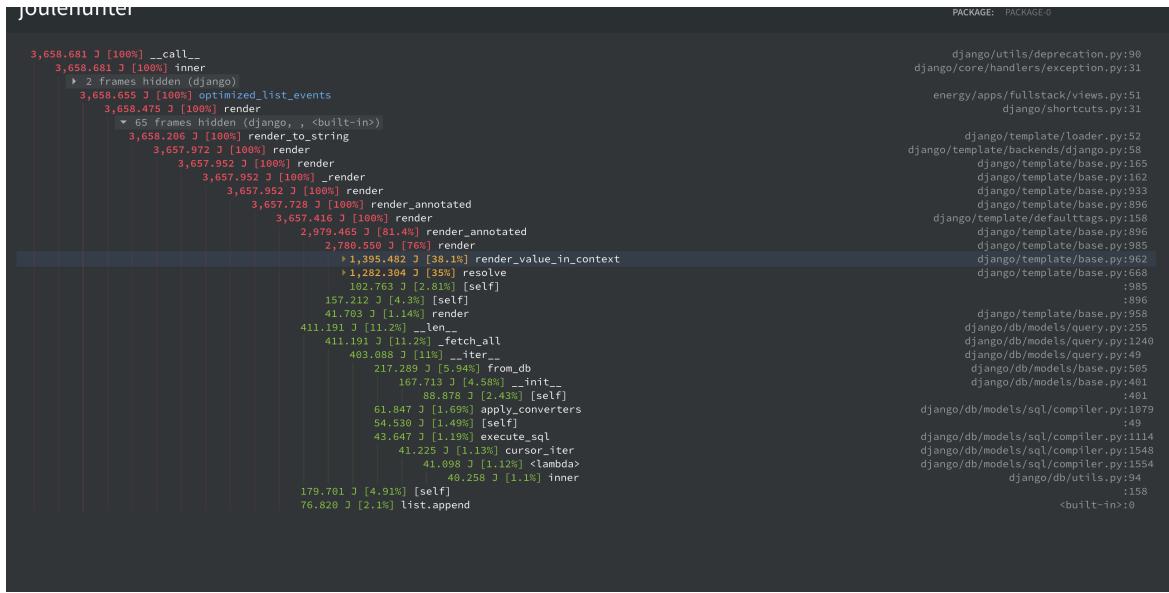


Figure 5.7: Tree representation of the energy consumption of a single request in django (naive version)

5.3.2 Python Insights

This section aims to formulate some actionable insights to optimize the energy consumption of Python applications without altering the source code. Therefore, we start by extending the work of Hasan et al. and (**author?**) to the context of Python environments.

Another field of investigation is the type of data and control structures that might impact the energy consumption. We iterate over a list using 3 methods. First, the classical `for(i in range(len(n))`). However, as we can see here, unlike other programming languages, it requires extra operations, such as determining the length of the collection and then using the iterator range. So, we tried the more adapted version `for(element in collection)`. Moreover, in most programming languages. the `for` loop is translated to a `while` loop (transformation from D type to B type – asm –), therefore we wanted to compare this with a `while` version. After determining the main ways to iterate over a loop, we run the collection algorithms of different data types to observe if they impact the energy consumption of the code, and we repeated it for the size of the collection.

As one can see in Figure 5.9, the type of the data has no impact on the energy consumption. However, the way one iterates over the collection has a huge factor. Interestingly, the `for in range` loop was by far the optimal one, followed by the regular `for in collection`, and the `while` part was the last one with an overhead of 400% compared to the first option.

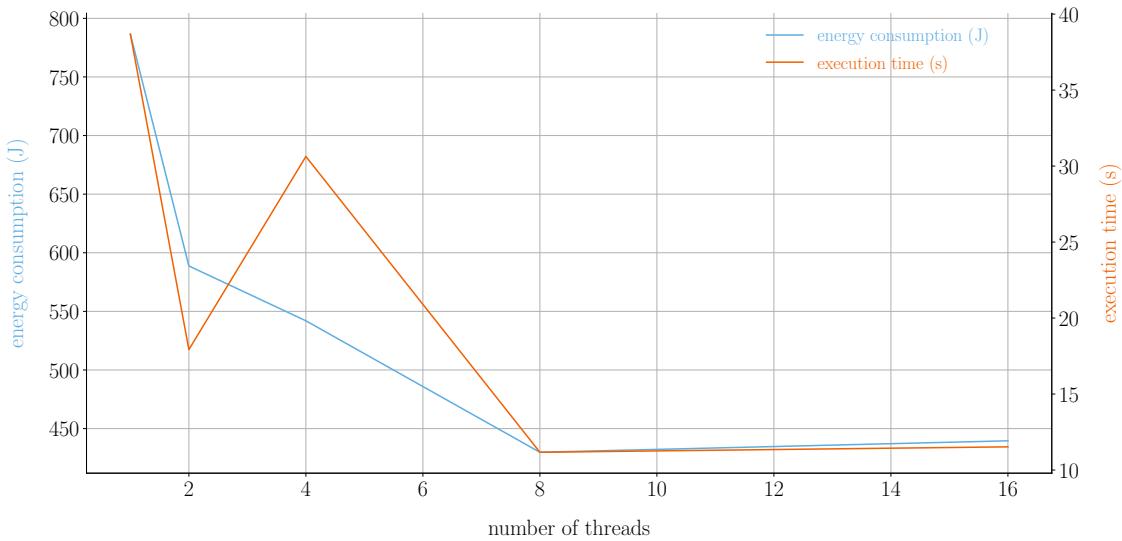


Figure 5.8: Energy consumption of Python multiprocessing depending on the number of exploited threads.

The reason behind such a behavior is mainly related to how the Python interpreter is implemented¹⁰ ¹¹. To reduce the latency of Python applications, most of the built-in functions and operations are written in C, and the same goes for the function `range`.¹² Furthermore, the function `len` has a complexity of $\mathcal{O}(1)$ as it is based on the function `Py_SIZE` of C, which stores the length in a field for the object¹³. Therefore, the `for` in `range` is creating a new iterator that has the same length as the first one and, for each iteration, requires a second access (`l[i]`) instead of one—explaining the doubled time. The `while` is even slower due to the implicit increment of the variable, which causes an extra operation during the loop. To confirm this hypothesis, we tried to construct a new list by editing the elements of the previous one (cf. Figure 5.10). And, as predicted, the built-in methods are the most energy-saving ones, while the customize `while` loop is the heaviest.

Another interesting finding is the impact of anonymous functions (also known as `lambda` expressions) on energy consumption. The reason is that Python treats these functions as local variables, unlike the predefined ones which are global in our case. Therefore, they are faster and consume less energy.

¹⁰<https://www.python.org/doc/essays/list2str>

¹¹<https://www.pythontutorial.net/python-basics/python-for-vs-while-loop/>

¹²<http://python-history.blogspot.com/2010/06/from-list-comprehensions-to-generator.html?m=1>

¹³<https://wiki.python.org/moin/TimeComplexity>

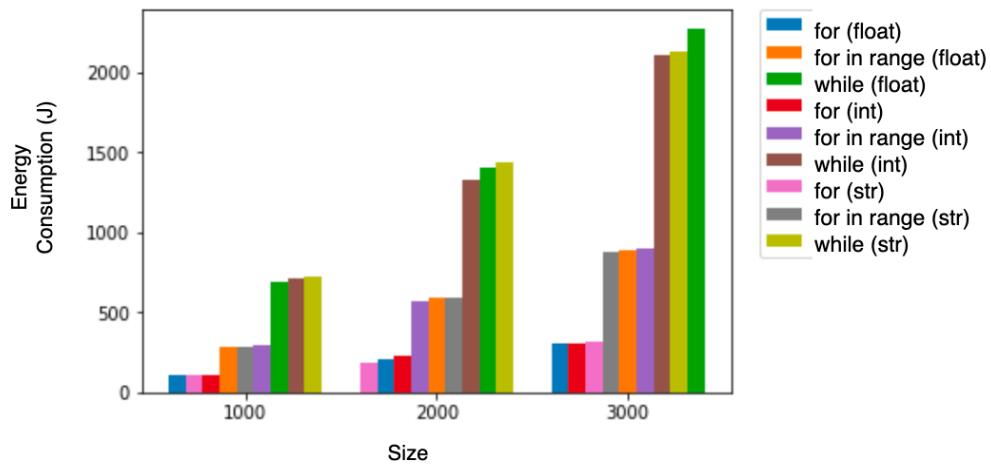


Figure 5.9: Comparison of the energy consumption of different Python loops.

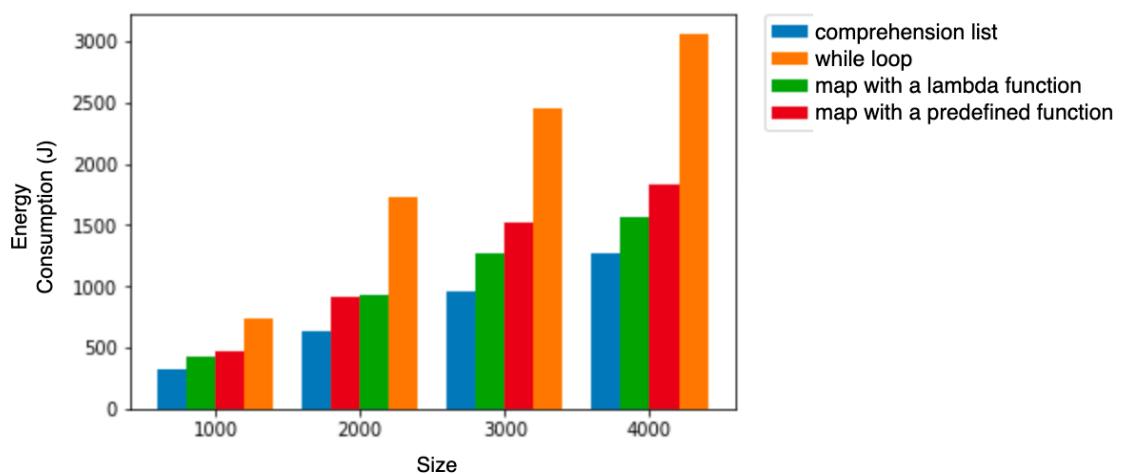


Figure 5.10: Comparison of the energy consumption of different methods to convert a list.

Conclusion This study demonstrates that the optimal way to reduce the energy consumption of Python application is to follow the guidelines and to privilege the built-in functions.

5.3.3 Python & Multiprocessing

Figure 5.11 compares the behaviour of python programs when we try to introduce the parallelism. As we know python is a single threaded program thanks to the GLS (Global Lock System) however due to the increase of the number of cores /threads per cpu it many libraries started to take advantage of this feature so most of them they will try to simulate the multithreading by using multiple instances of the processor

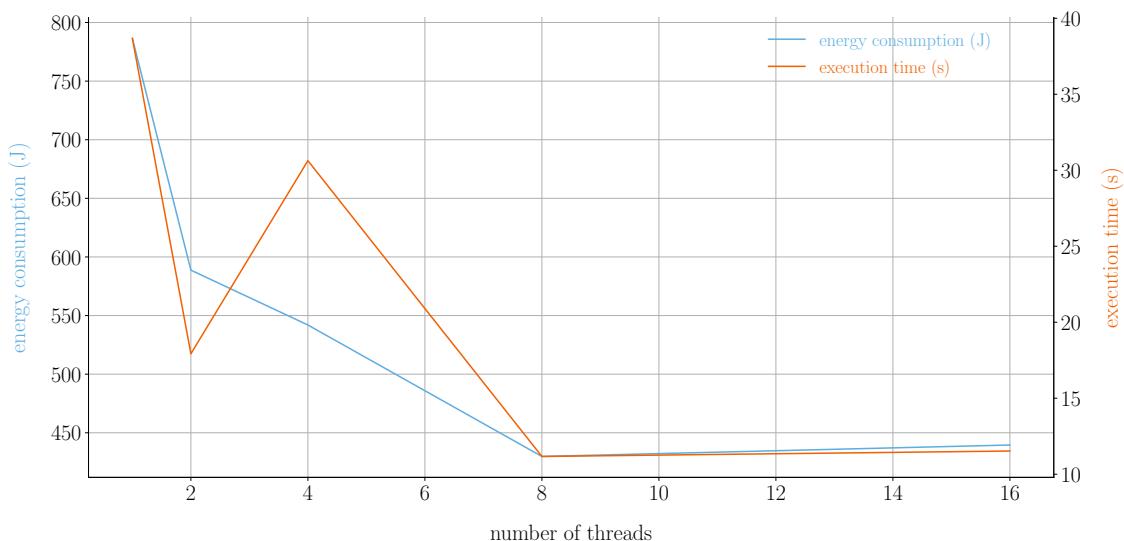


Figure 5.11: energy behaviour based on multiprocessing

As we can see in the graph the energy consumption is correlated with the number of threads until we arrive to the limit of the cores and then we lose the advantage of the multiprocessing, well in that case we pass from parallelism to concurrency. where different sub processes have to compete for the CPU Resources. Another finding is when we hit the number of physical cores. there was an increase of execution time but still reduced energy, the reason behind this is the scheduler of the operating system. basically he favors the hyper threads than the physical core which will lead to some context switches which cause the slow behaviour, however in the other hand the other two physical cores are not consuming energy, neither their hyper threads which explains the gain of the energy consumption in this case. in the Chapter we discuss a deeper this behaviour of the scheduler and in that case we confirm that it is not related to python but it is more generic behaviour

5.3.4 Python & Machine Learning

Machine learning is becoming an integral part of our daily lives, growing more potent and energy-hungry each year.

As machine learning can have a significant impact on climate change, it is vital to investigate mitigation techniques.

Hardware settings Chifflet 8 from Grid 5000's Lille site was used for all of the trials. The machine is outfitted with two Intel Xeon Gold 6126 CPUs, each having 12 physical cores, 192 GB of RAM, and two 32 GB Tesla V100 GPUs.

Software settings For the sake of reproducibility, each experiment is done within a Docker container using Jupyter lab. These tests are run on top of a minimal version of Debian-10 to increase the accuracy of the tests by eliminating any unnecessary processes.

Input Workload

Models Several models were developed, however, only two were used in the final trials because they achieved 94 percent accuracy in an acceptable length of time. David Page's cifar10-fast¹⁴ and Woonhyuk Baek's torch skeleton¹⁵ are shown here.

Datasets The CIFAR-10 dataset was the major source of data for the studies. It is made up of 60000 32x32 color images grouped into ten categories.

Some experiments were done using the MNIST dataset of handwritten digits to validate the results acquired from the first dataset. The model did not need to be updated because the 60000 28x28 grayscale photos were padded.

candidates

The experiments were run with several different CPU and GPU configurations:

- with and without GPU,
- with and without CPU hyper-threading,
- different number of CPU physical cores.

¹⁴<https://github.com/davidcpage/cifar10-fast>

¹⁵<https://github.com/wbaek/torchskeleton>

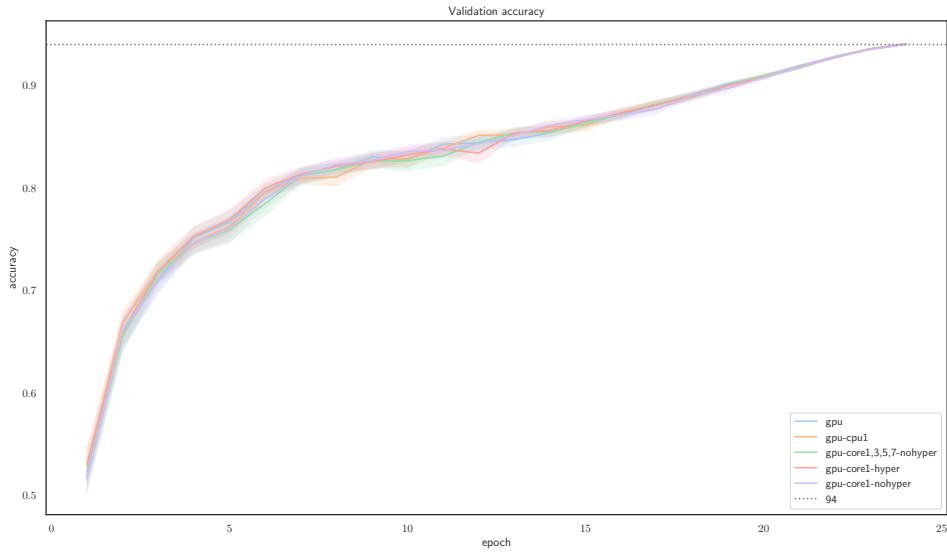


Figure 5.12: accuracy based on epoch

Key Performance Metrics

We used Pytorch 1.10.0 to train those models and pyjoules to measure the energy consumption of the GPU and CPU.

- accuracy : in %
- execution time : in seconds for both the duration of each epoch and the total duration to achieve a certain accuracy
- total energy consumption : for the CPU and the GPU
-

5.4 Results & Findings

As the model's accuracy increases, so does the energy required for the next accuracy increment.

Figure 5.13 depicts how the curve steepens as training progresses. For example, training to 90% accuracy requires three times the energy required for training to 80% accuracy.

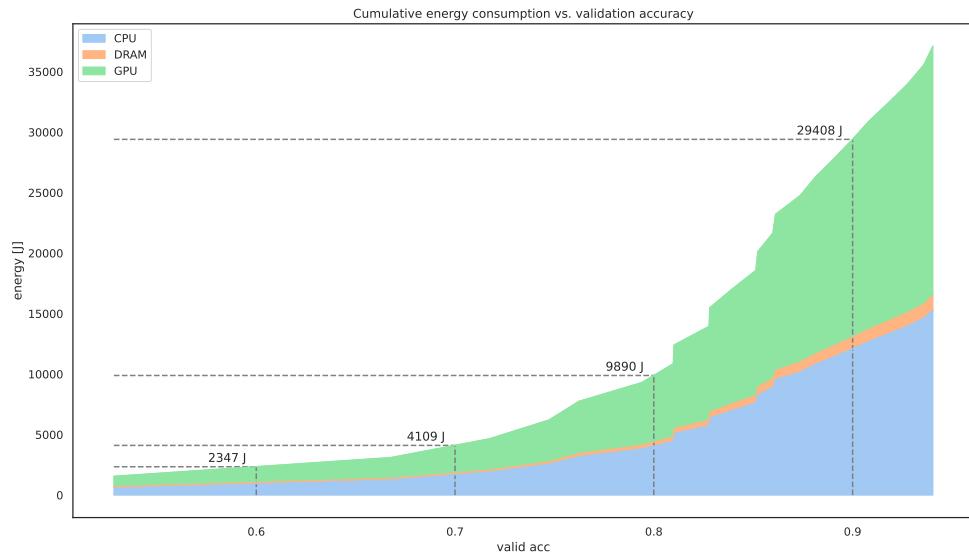


Figure 5.13: cumulative energy consumption vs accuracy

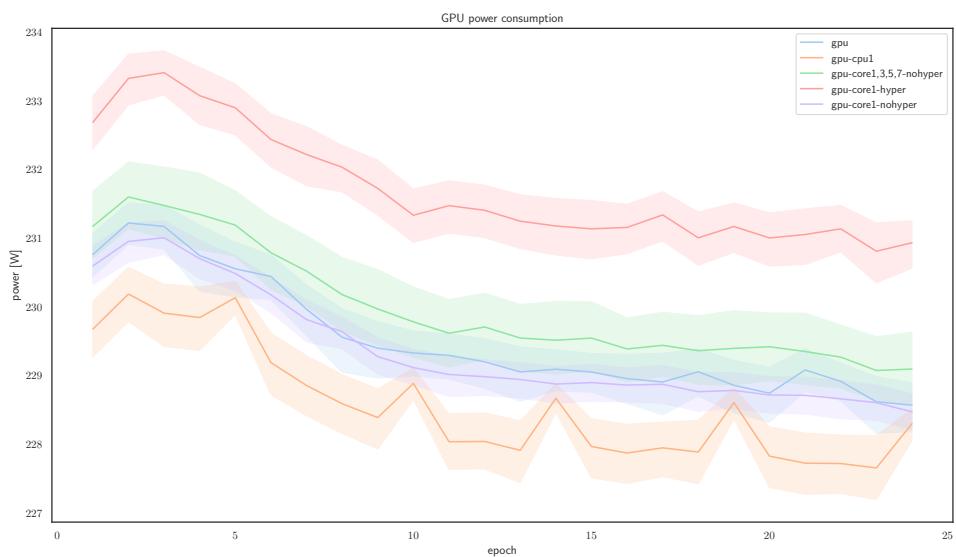


Figure 5.14: evolution of average gpu power based on epoch

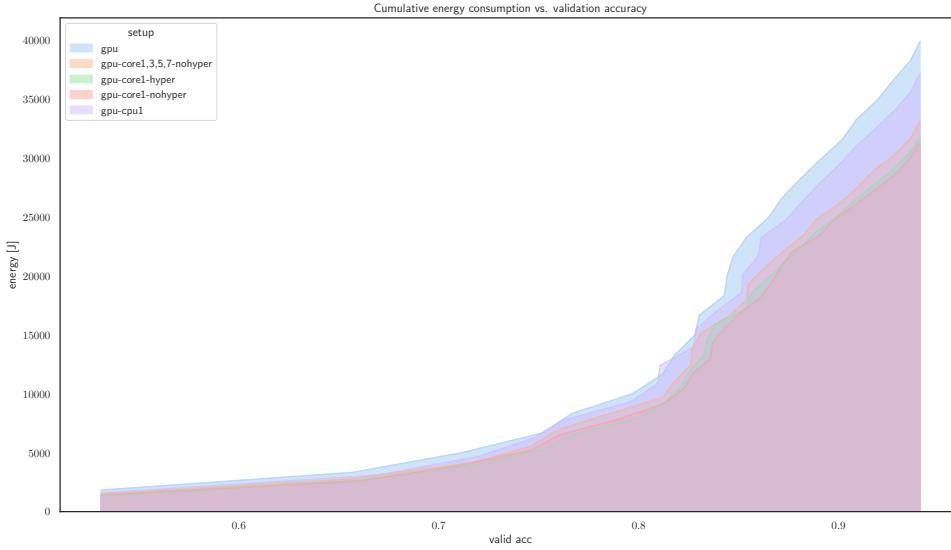


Figure 5.15: cumulative energy of fast10 benchmark within different configurations

Conclusion We discovered that when a model's accuracy improves, so does the energy required for a subsequent accuracy gain. This begs the question of when we should discontinue training. Is a 10% increase in accuracy worth it if we have to spend three times the energy? Even while utilizing a GPU consumes more total power, the training time is significantly reduced. As a result, the utilization of a GPU is required to lower the energy consumption of the training. Some experiments revealed unusual behavior that was not studied during this internship. These could be examined in future studies and perhaps used to achieve more energy savings. The execution time did not depend on the number of cores employed in several studies. This parameter had a significant impact on training time in others. As a result, the best core configuration will be determined by the script. The next steps in this area could include exploring the reasons for these discrepancies and possibly creating a script that can automatically find the ideal core configuration for a specific script.

5.5 Python Interpreters

Due to the lack of support for most non-conventional Python interpreters, we mainly focus on micro-benchmarks. Except for PyPy, most of the Python implementations do not support extra Python libraries, despite those extra implementations being developed to optimize a specific library, such as Numba with Numpy, or intelpython with machine learning algorithms.

Preliminary studies For the first studies, we used the official version of Python, because the goal was mainly to highlight the impact of the structure of code on energy consumption. One main drawback of the previous method is the work to be done to update the existing code base to reduce energy consumption. To avoid such hustle, we tried to find a non-intrusive approach to make the Python code more eco-friendly without altering its structure. Python is an interpreted language, which led many initiatives to implement their own interpreter to improve one or many aspects of the Python code. In the following section, we discuss the impact of those implementations on the energy consumption of python programs, and in which case, one should use a non-conventional interpreter to save the energy consumption of their application.

To do so, we gathered a list of interpreters, transpilers and other optimization libraries that can contribute to reduce the energy consumption of legacy Python applications:

1. **CPython:**¹⁶ This Python interpreter, written in C, is the reference interpreter of Python. CPython compiles the source code into byte-code and then interprets it. The CPython project supports both versions of Python 2 and 3;
2. **PyPy:**¹⁷ An alternative implementation of the Python interpreter. It is written using *RPython* to use the JIT. It compiles the most used portions of the Python code into a binary code for better performance. To benefit from these optimizations, the program has to be executed for at least for few seconds so the JIT has enough time to warm up, the JIT optimization are only applied to the code written by the developer and not to external libraries;
3. **Cython:**¹⁸ A static compiler for Python. It translates the Python code into C, and then compiles it using a C compiler. It also supports an extended version of the Python language that allows programmers to call *C functions*, declare *C types* and use static types, which will help the translation of Python objects into native types, such as integers, float. This often means better performances, since native C libraries are almost all the time faster than the Python written once [107];
4. **Intel Python:**¹⁹ A customized interpreter developed by Intel to enhance performances of Python programs. It is dedicated to data sciences and high-performance computing.

¹⁶<https://www.python.org/>

¹⁷<http://pypy.org>

¹⁸<https://github.com/cython/cython>

¹⁹<https://software.intel.com/en-us/distribution-for-python>

It uses some Intel kernel libraries, such as Math Kernel Library (Intel MKL²⁰) and data analytics acceleration library (Intel DAAL²¹). It supports both versions of Python;

5. **Active Python:**²² It is developed by the Activestates company and provides a standardized Python distribution to ensure license compliance, security, compatibility and performance. Therefore, ActivePython implements its built-in packages (more than 300 packages) and supports both versions of Python;
6. **IronPython:**²³ A .Net-based Python interpretation platform written in C# that is used with the .Net virtual machine or Mono. It benefits from all the optimizations of .Net virtual machines, such as the JIT and garbage collector mechanisms;
7. **GraalPython:**²⁴ A Python interpreter that is based on GraalVM²⁵ (a universal virtual machine developed by oracle for running applications written in different programming languages). For the time being, it only supports Python 3 and it is still in the experimental stage;
8. **Jython:**²⁶ An implementation of Python programming language written in Java for the *Java Virtual Machine* (JVM). Similar to IronPython and GraalPython, it leverages the optimization mechanisms provided by the JVM to enhance the Python performances;
9. **MicroPython:**²⁷ A lightweight Python version dedicated to embedded systems and micro-controllers;
10. **Nuitka:**²⁸ A Python compiler written in Python that generates a binary executable from Python code. It translates the Python code into a C program that is then compiled into a binary executable;
11. **Numba:**²⁹ A library that includes JIT compiler to enhance the performances of Python functions using the industry-standard LLVM compiler library;

²⁰<https://software.intel.com/en-us/mkl>

²¹<https://software.intel.com/en-us/intel-daal>

²²<https://www.activestate.com/products/activepython/>

²³<https://ironpython.net>

²⁴<https://github.com/graalvm/graalpython/>

²⁵<https://www.graalvm.org/docs/why-graal/>

²⁶<https://jython.github.io>

²⁷<http://micropython.org>

²⁸<http://nuitka.net/pages/overview.html>

²⁹<https://numba.pydata.org>

12. **Shedskin**:³⁰ A static transpiler that translates implicitly statically typed python into C++ code;
13. **Hope** [4]: A Python library that aims to introduce JIT compiler into the Python code;
14. **Parakeet** [?]: A runtime accelerator for an array-oriented subset of Python;
15. **Stackless Python**:³¹ An interpreter that focuses on enhancing multi-threading programming;
16. **Pyjion**:³² A JIT API for CPython, same purpose as Parakeet and Hope;
17. **Pyston**:³³ A performance-oriented Python implementation built using LLVM and modern JIT techniques. The project is funded by Dropbox;
18. **Grumpy**:³⁴ A source-to-source transpiler that translates the Python code into Go before being compiled to a binary executable. It also offers an interpreter, called *grumprun*, which can directly execute the Python code. Unfortunately, we cannot use it because the project is already outdated (last commit is in 2017) and it has a lot of limitations in terms of supporting the Python language, such as some built-in functions and standard libraries;
19. **Psyco**:³⁵ A JIT compiler for Python;
20. **Unladen Swallow**:³⁶ An attempt to (use) LLVM as a JIT compiler for CPython.

5.5.1 Runtime Classification

Before further proceeding with the list of candidate runtime for Python applications, we propose a classification according to several criteria:

Type refers to the category of runtime infrastructure that supports the execution of a Python application. In particular, we consider 3 types of environments: *Interpreter*, *Compiler* and *Library*; *Interpreter* refers to the class of environment that does not require any preprocessing of Python source code; *Compiler* introduces a compilation phase before

³⁰<https://github.com/shedskin/shedskin>

³¹<https://github.com/stackless-dev/stackless/wiki>

³²<https://github.com/microsoft/pyjion>

³³<https://blog.pyston.org>

³⁴<https://github.com/google/grumpy>

³⁵<http://psyco.sourceforge.net>

³⁶<https://unladen-swallow.readthedocs.io/en/latest/>

the execution of the application. Finally, *Library* requires some modification of the source code;

Runtime refers to the technology supporting the execution of a Python application. This technology can refer to the programming language used to program the interpreter, the target language for a compiler or a library;

JIT optimization refers to the support of *just-in-time* compilation in the runtime infrastructure supporting the execution of the application;

GC optimization refers to the support of *garbage collection* in the runtime infrastructure supporting the execution of the application;

Python version(s) refers to the list of Python source code versions supported by the runtime environment.

We should explain the classification of these runtimes

There are other implementations that we did not consider because either the project aborted many years ago or it has very limited support for Python features. After the collection of those implementations, we filtered them. To keep only the versions that are still maintained and support most Python features. and we classified them into 3 categories depending on their integration with the python code. In Table 5.3, we describe the implementations that we kept and the version of each implementation and its category.

5.6 experimental protocole

As we have discussed in the previous chapter 3. instead of running the tests, the idea was to design a system that allows practitioners to reproduce and extends our tests. and then we use the same system to run answer some of researchs question.

5.6.1 measurement context

Hardware settings all our tests have been executed in a Dell PowerEdge C6420 server machine. A summary of its hardware is listed in table 5.4. The machine is equipped with a minimal version of Debian 9 (4.9.0 kernel version) where we install Docker (version 18.09.5).

software settings For the sake of reproducibility, each experiment runs within a Docker container.

Table 5.2: Classification of Python implementations

Name	Type	Runtime	Optimisations		Python	
			JIT	GC	2	3
CPython	Interpreter	C	—	—	✓	✓
Intel Python	Interpreter	C	—	—	✓	✓
ActivePython	Interpreter	C	—	✓	✓	✓
PyPy	Interpreter	Python	✓	✓	✓	✓
IronPython	Interpreter	.Net	✓	✓	✓	✓
GraalPython	Interpreter	GraalVM	✓	✓	—	✓
Jython	Interpreter	Java	✓	✓	✓	—
Stackless Python	Interpreter	Python	—	—	✓	—
MicroPython	Interpreter	c	—	—	—	✓
Pyston	Interpreter	LLVM	✓	—	✓	—
Unladen Swallow	Interpreter	LLVM	✓	—	✓	—
Cython	Compiler	C	—	—	✓	✓
Nuitka	Compiler	C	—	—	✓	✓
Shedskin	Compiler	C++	—	—	✓	✓
Grumpy	Compiler	Go	—	—	✓	✓
Numba	Library	C	✓	—	✓	✓
Hope	Library	Python	✓	—	✓	✓
Psyco	Library	Python	✓	—	✓	✓
Pyjion	Library	.NET Core	✓	—	✓	✓
Parakeet	Library	C	-	—	✓	—

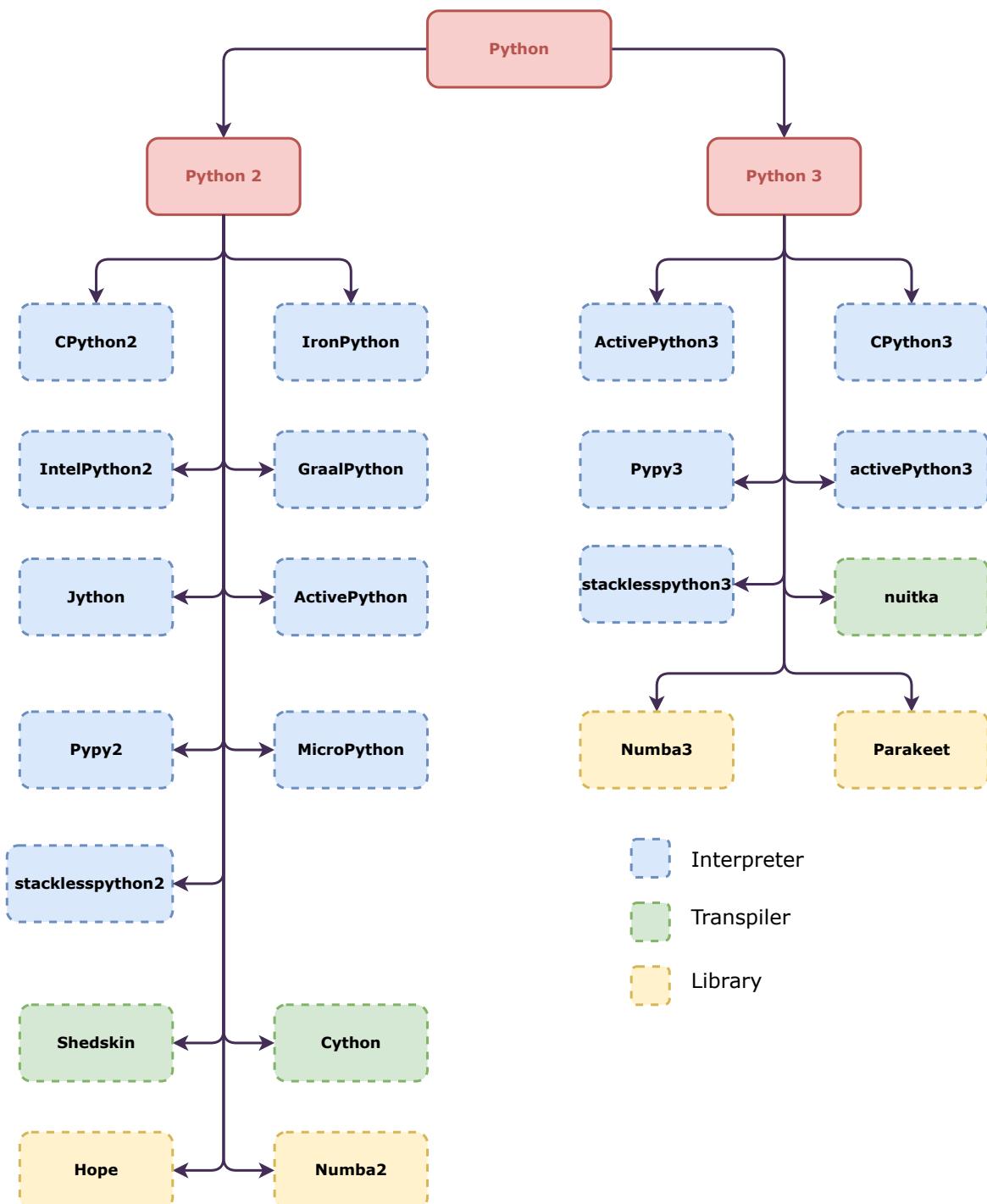


Figure 5.16: Python interpreters

Table 5.3: Classification of Python implementations

Version	Interpreter	Transpiler/Compiler	Jit library
Python 2	Cpython2 Pypy2 Pyton Ironpython Jython Micropython Pysec StacklessPython	Cython2 Shesdskin Grumpy	Numba 2 Hope Parakeet Psyco Pyjion
Python 3	Cpython3 Pypy3 GraalPython	Nuitka	Numba3

CPU	Intel Xeon Gold 6130 (Skylake, 2.10GHz, 2 CPUs/node, 16 cores/CPU)
Memory	192 GiB
Storage	240 GB SSD SATA Samsung MZ7KM240HMHQ0D3 480 GB SSD SATA Samsung MZ7KM480HMHQ0D3 4.0 TB HDD SATA Seagate
Network	eth0/epn24s0f0, Ethernet, configured rate: 10 Gbps, model: Intel Ethernet Controller X710 for ib0, Omni-Path, configured rate: 100 Gbps, model: Intel Omni-Path HFI Silicon 100 Series [d]

Table 5.4: Testing Machine Configuration

5.6.2 Metrics

Our focus will be mainly on CPU energy consumption because it is ten folds more than the DRAM one, since it is finit job benchmarking, time is highly correlated within the energy, and it will be only useful to explain certain energytical behaviour so we wont put a lot of focus on this metric.

Energy measurement As we know, the energy of a program is the integrale of its power overtime. For us ower case, we used Intel *Running Power Average Limit* (RAPL) [68] to collect the power samples of the running tests,

We used used POWERAPI [32], to report Data collected by intel Rapl and send it into another machine that we call computing machine. then we calculate the Energy using the the trapezoidal rule.

5.18

figure 5.17 shows the architecture of our testing model.

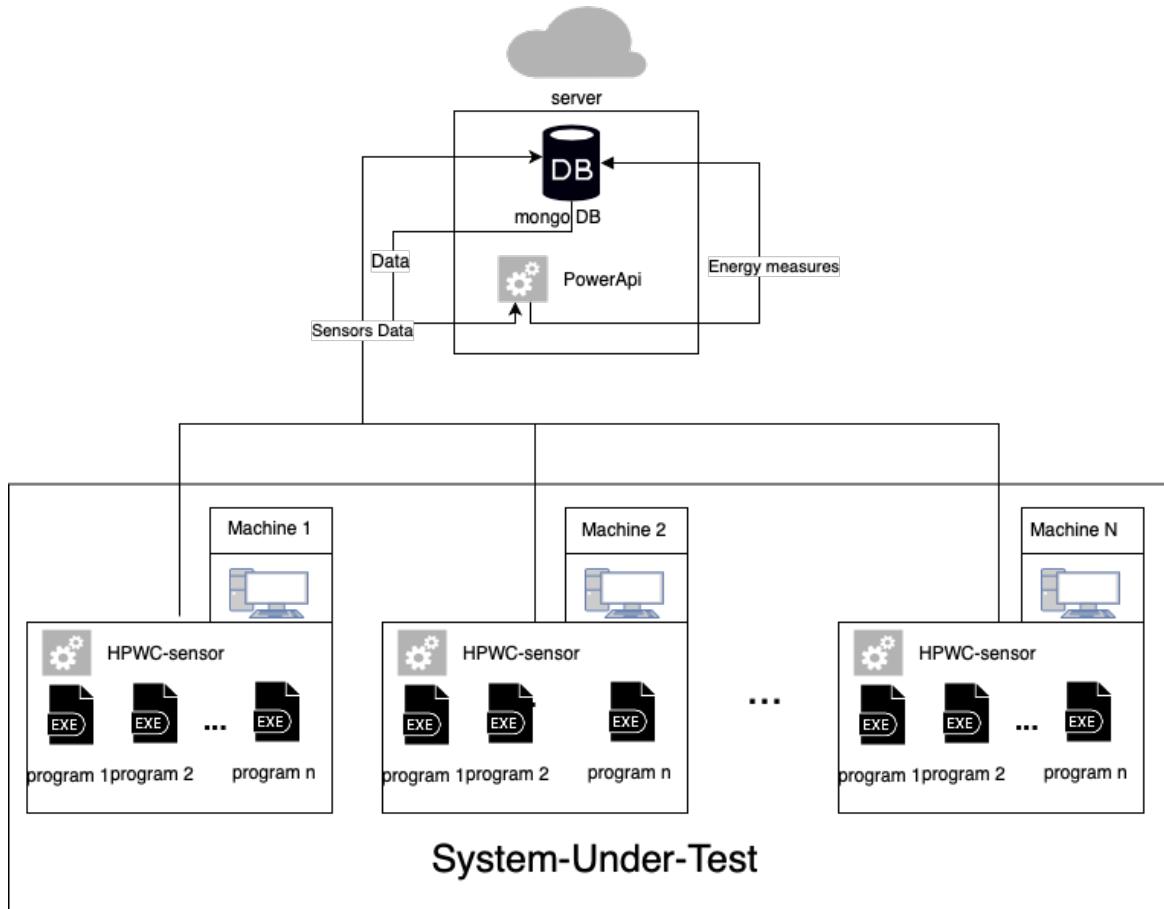


Figure 5.17: powerapi architecture

The reason of separation between data collection and energy calculation is to minimise any interference with the test so our sensor is a light c program running inside a docker container.

$$E = \int_b^a P(t)dt \simeq \sum_{k=1}^n \frac{P(t_k - 1) + P(t_k)}{2} \quad (5.1)$$

Figure 5.18

5.6.3 tests preparation

workload To test our implementations, we employed the Tommti microbenchmarks suite³⁷. A set of 13 microbenchmarks that examine common language features such as arithmetic operations, data structures and input/output manipulations, among others. In addition to these tests ,we implemented some binary operations to examine the behavior of the previous implementations when it comes to low-level operations that works only with registries.

To study the behaviour of the python implementation regarding the energy consumption, we have to focus on the effect of the implementation and mitigate the maximum any side effects such as the organisation of the code or any extra consumption due to the operating system or tier libraries. Therefore, for each test we took the implementation written in python2 as a reference and tried to use it in other implementations as it is. If it is not supported by python3, we transformed the code using the officiel library *2to3*³⁸. In the case of the libraries that use *JIT* adding a decorator to the function that we want to optimize was enough, if there are other changes we assume that they alter the original code which is against our purpose.

Each test is implemented in a Docker container for the several reasons

- Isolation: each container has only the test program implemented with a single python runtime to remove any interference between different implementations.
- Deployment: to use the testing machine without extra configurations that may alter the behaviour of the os toward the enrgy consumtions
- Reproducibility: One of the most frequent benchmark crimes [118] in research is the lack of Reproducibility, by using Docker we ensure that each test has an Image that will be accessible in public.

Despite the presence of the official docker images for the most of the runtimes, we prefered to build our own using the same reference image in order to remove any bias due to the Os used in the offical image. We used ArchLinux with the kernel version 4.9.184 as a base image.

5.6.4 Extension

As we have done with the previous chapters. we provide a tool that allows to extend the tests with new workloads and new candidates. In the repository³⁹. we have a tool that allows to

³⁷<http://www.tommti-systems.de/main-Dateien/reviews/languages/benchmarks.html>

³⁸<https://docs.python.org/3.7/library/2to3.html>

³⁹<https://github.com/chakib-belgaid/python-implementations>

generate new workloads and new candidates. The script `generator.py` allows to create new benchmarks by implementing a python code within different interpreters. Then it generates `launcher-benchmark.sh` that can be used directly to run the experiment. Furthermore all the successful implementations are stored in separate directory and the that couldn't work (mostly because of compatibility issues) stored in a recap file called `benchmarkTest.md`. where benchmark is the name of the new workload; To add extra **candidates** one should add a base docker file that contains the new implementation and if there should be extra manipulation that should be added to the workload files ,such as adding new decorator or changing some parameters, then they should be added as an extra function in the script `generator.py`. Finally they should be included in the python candidates.

5.7 Results and finding

This section will be dedicated to the results of the experiments and a statistical analysis of these results.

Because we are conducting a comparison study, we first did a Shapiro-Wilk [114] test to see if the data follows a normal distribution. Despite, the fact some implementations had a p-value bigger than $a=0.05$ (such as cpython and active python), other implementations like pypy and numba presented a p-values smaller than 0.01 which leads to reject the hypothesis H_0 that all the distributions follow a normal distribution. Therefore,to compare the different implementations, we performed the non-parametric Mann-Whitney U test [129], with CPython2 and CPython3 as base references.

The Mann-Whitney U test results for each implementation are shown in Table 5.5. The first column displays the implementation's average energy usage, while the second provides the p-value of the test when compared to the reference Python implementation. If the p-value is less than 0.05, the test result is significant. This implies that the given implementation's energy consumption differs significantly from the reference implementation's. As one can notice, most of the implementations differ significantly from the reference implementation.

We want to study the position of these implementations toward the references. To do so, we proceed with an agglomerative hierarchical cluster analysis (HCA) [66]. The HCA is a method that allows grouping the different implementations based on their energy consumption. This technique is a bottom-up approach that starts with each implementation as a single cluster and then merges the two most similar clusters until all the clusters are merged into a single cluster. The similarity between two clusters is measured by the maximum distance between their furthest points. The distance between two points is calculated as the

Implementation	array		intArithmetic		doubleArithmetic		hashes		heapsort		trig	
activepython	energy (J)	<i>p-values</i>	energy (J)	<i>p-values</i>	energy (J)	<i>p-values</i>	energy (J)	<i>p-values</i>	energy (J)	<i>p-values</i>	energy (J)	<i>p-values</i>
activepython	4e02	0.008	6.9e02	0.002	6.6e02	0.002	9.7e02	0.008	2.9e02	0.008	5.1e02	0.008
cpython2	3.4e02	base	5.6e02	base	5.5e02	base	6.5e02	base	2.9e02	base	4.1e02	base
cpython3	3.2e02	base	7.4e02	base	7.3e02	base	7.9e02	base	2.4e02	base	4.1e02	base
graalpython	1.2e05	0.008	24	0.002	24	0.002	6.4e02	0.008	1.3e05	0.008	72	0.008
intelpython2	3.7e02	0.690	5.9e02	0.015	5.6e02	0.485	7.1e02	0.008	2.7e02	0.690	4.2e02	0.151
intelpython3	3.5e02	0.008	7.7e02	0.002	7.6e02	0.026	9.4e02	0.008	2.6e02	0.008	4.8e02	0.008
ipy	3e02	0.008	4.4e02	0.002	4.7e02	0.004	1.3e03	0.008	2.6e02	0.016	4.5e02	0.008
jython	5.2e02	0.008	1.3e02	0.002	1.6e02	0.002	6.4e02	0.056	4.6e02	0.008	6.2e02	0.008
micropython	3.1e02	0.008	8.2e02	0.002	8.4e02	0.004	9.5e03	0.008	3.4e02	0.008	5.3e02	0.008
nuitka	2.9e02	0.008	5.4e02	0.002	5.5e02	0.004	9.5e02	0.008	2.2e02	0.008	3.9e02	0.008
numba2	1.9e02	0.008	27	0.002	36	0.004	6.8e02	0.008	2.1e03	0.008	4.4e02	0.008
numba3	11	0.008	10	0.002	10	0.004	9.1e02	0.008	7.2e02	0.008	4.5e02	0.008
pypy2	16	0.008	29	0.002	30	0.002	1.1e02	0.008	20	0.008	64	0.008
pypy3	13	0.008	17	0.002	18	0.004	1.9e02	0.008	20	0.008	65	0.008
shedskin	47	0.008	7.4	0.002	7.3	0.004	1.1e03	0.008	44	0.008	7.7	0.008
Implementation	longArithmetic		matrixMultiply		io		stringConcat		nestedLoop		except	
activepython	energy (J)	<i>p-values</i>	energy (J)	<i>p-values</i>	energy (J)	<i>p-values</i>	energy (J)	<i>p-values</i>	energy (J)	<i>p-values</i>	energy (J)	<i>p-values</i>
activepython	6.7e02	0.008	4.3e02	0.008	2.1e02	0.016	14	0.310	4.1e02	0.008	2.6e02	0.008
cpython2	5.5e02	base	4e02	base	2e02	base	12	base	4.2e02	base	4.3e02	base
cpython3	7.4e02	base	4.4e02	base	2e02	base	13	base	3.8e02	base	2.3e02	base
graalpython	24	0.008	45	0.008	7.5e02	0.008	26	0.008	12	0.008	1.6e02	0.008
intelpython2	5.7e02	0.016	4.8e02	0.008	2e02	0.421	13	0.151	4.4e02	0.008	4.6e02	0.008
intelpython3	7.7e02	0.008	5e02	0.008	2.2e02	0.008	14	0.310	4.4e02	0.008	2.7e02	0.008
ipy	4.7e02	0.008	4.1e02	0.008	3.8e02	0.008	49	0.008	3.3e02	0.008	7.2e02	0.008
jython	1.6e02	0.008	1.9e02	0.008	2e02	0.310	21	0.008	1.9e02	0.008	7.1e02	0.008
micropython	8.3e02	0.008	5.3e02	0.008	5.3e03	0.008	44	0.008	4.3e02	0.008	3.6e02	0.008
nuitka	5.4e02	0.008	4.3e02	0.421	2e02	0.151	12	0.310	3.6e02	0.008	2.2e02	0.095
numba2	35	0.008	4e02	0.310	2.2e02	0.008	20	0.008	14	0.008	4.5e02	0.008
numba3	11	0.008	4.2e02	0.222	2.1e02	0.032	19	0.008	10	0.008	2.4e02	0.008
pypy2	30	0.008	24	0.008	1.7e02	0.008	7.7	0.008	27	0.008	14	0.008
pypy3	17	0.008	23	0.008	2.6e02	0.008	8.3	0.008	23	0.008	14	0.008
shedskin	7.8	0.008			3.8e02	0.008	44	0.008	7.1	0.008	5.6e02	0.008

Table 5.5: Energy consumption python environment

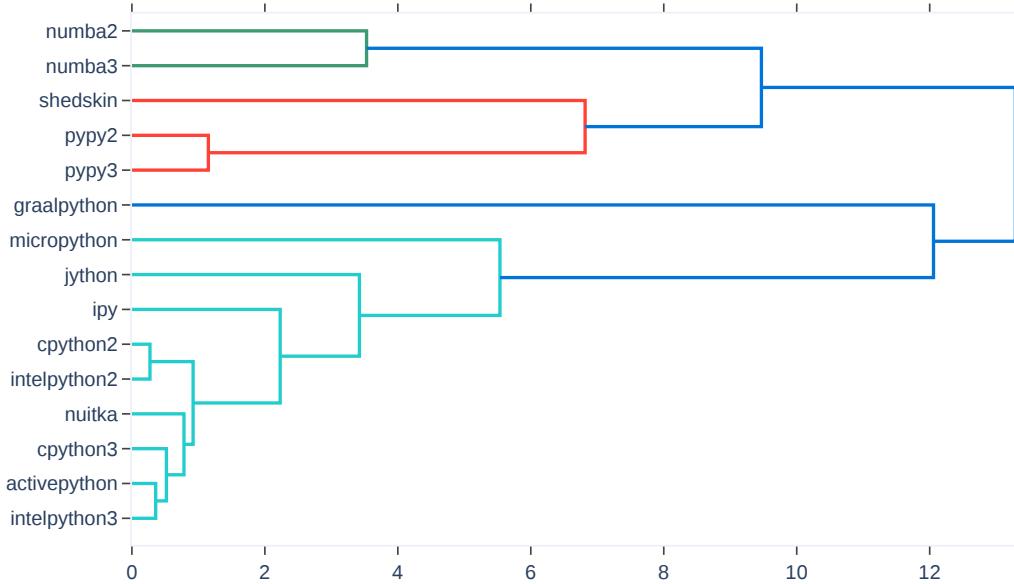


Figure 5.19: Dendrogram of the different implementations

Euclidean distance between the two points. Figure 5.19 displays the dendrogram produced by the HCA

Except GraalPython, one can notice 3 main clusters. The first cluster contains the reference implementations (CPython2 and CPython3) and the implementations that are based on interpreters (Ironpython,Intelpython, ActivePython, Jython, and Nuitka). Moreover, each implementation is the closest to its reference Python version. The interpreters that are based on other virtual machines, such as Jython with JVM and IronPython with.NET, behave slightly differently from the reference implementation. As for the micropython, It is the furthest from the references due to his behavior toward exceptions. The second cluster contains the interpreter implementations that are based on a JIT compiler (Pypy2 and Pypy3). Closer to them we find shedskin, which is a transaltor that converts the Python code into C++ code to be later compiled into binary code. The results table 5.5 shows that this cluster is by far the greenest one. While numba2 and numba3 are also JIT libraries, they differ from PyPy due to their manual optimisation. Unlike pypy, which will decide by itself which part of Python to optimize. In Numba, the developer must target the portion of the code that will

be passed to JIT, while the rest of the code will be executed by the regular interpreter (aka CPython).

Remark : Because graalPython was still in its early stages, some techniques required an abnormally lengthy time, affecting the clustering Algorithm.

To help practitioners to choose the best implementation for their use case, we have created a chart that describes the behavior of each implementation toward a specific aspect of programming. Figure 5.20 present a radar plot for all the implementations that have been used in this experimentation.

This graph depicts each implementation's energy score in relation to each benchmark. The lower the energy consumption, the higher the score. We adopted a logarithmic scale to allow the reader to compare numerous implementations due to the large differences between them.

As one can notice in figure 5.20, there is no evolution between CPython2 and CPython3. Moreover, Intelpython, and Active Python all behave similarly to the reference implementation . Therefore, one can conclude that the work done on those interpreters is primarily to improve a specific purpose and not the core interpreter. Active Python states that their version is focused on security and prepackaged libraries, which explains why it is slower than other versions due to the addition of more security reinforcements⁴⁰ . As for IntelPython,it is designed for machine learning. Unfortunately, the Tommi benchmark is geared for general-purpose programming. Despite the fact that Nuitka is a compiler, there was no discernible difference in energy consumption. It was even more similar to the reference Python implementation than other interpreters. However, if we look at the nuitka techniques, we can see that they simply encapsulate the Python code with an interpreter into a single executable. However, shedskin exhibit has the best energy consumption pattern when it comes to arithmetic operations. One can conclude it is due to the fact of the native type of the variables. Unlike in the JIT, where they are treated as objects in the beginning.

for the other VM-based interpreters. Jython and Ipy lacked in terms of energy optimisation, which was kind of expected since they were at the beginning of the development stage and the main purpose of such implementation is to link the bytecode generated by Jython and IronPython with their respective virtual machines.

Unlike the previous interpreters, GraalPython exhibits a certain promise when it comes to complex algorithms (nested loops). Finally,micro python is dedicated to embedded systems, so claiming it is a powerful CPU machine will be misleading.

⁴⁰<https://www.activestate.com/solutions/why-activestate/>

When it comes to the input outputs operations , when it came to input/output manipulation. Except for Jython,most of the interpreters behaved similarly . Because these operations are more dependent on the hardware than the interpreter.

5.7.1 python and binarry operations

The second part of this study, is to look at how different Python implementations work while treating different binary operations .

Table 5.21 provides a detailed report on the energy consumption of different implementations of the code. Besides *Cython*, each implementation treats the binary operation the same. As one can see, Table 5.21 shows no difference in the energy consumption of those operations.

However, when it comes to Cython, the rotation operations tend to consume 5 times more energy than the other operations, and left rotations consume even more than the right ones.

The reason behind such behavior is the fact that Cython does not have a native implementation for these two operations, so it replaces them with more complex operations, which leads to an extra time of 6s for the right rotation and 10s for the left, compared to 1.5 for the other operations. -

Therefore, it is safe to summarize all the operations into one that we call *binary*.

Figure 5.22 shows the average energy consumption of different implementations for the binary operations. One can notice 3 clusters.

The first, which is the interpreted one, uses approximately 500 joules per operation. The second is a non-optimized compilation that includes Cython and Nuitka and uses 10 times less energy than the first. And the final one, which has been fully optimized, uses 50 percent less energy. This category includes jit libraries, also known as numba, Python implementations that include JIT, also known as pypy, and shedskin, which transpile the source code into C and then use a C/C++ compiler to generate the binary, as opposed to cython and nuitka, which directly compile the python code. To conclude this. JIT has a huge impact when it comes to binary operations, and the wrong choice of the implementation can lead to up to a 100x increase in energy consumption. In the other hand,Different binary operations exhibit the same behaviour when it comes to the energy consumption for all the implementations except Cython.

5.8 conclusion

One can conclude that the choice of Python interpreter has a significant impact on the programs' energy consumption. This investigation is made more intriguing by the absence of a universal solution. The primary downside is the incompatibility of some of these solutions, which causes us to make concessions when we need a generic answer.



benchmark	Addition	Right Rotation	Left Rotation	Or	XOR	Average
activepython	676.980	763.208	651.783	743.016	728.828	712.76
cpython2	441.082	435.886	430.846	415.247	419.081	428.43
cpython3	595.209	685.085	563.839	657.972	655.560	531.53
cython	35.077	182.688	274.177	34.868	34.504	112.26
nuitka	33.260	32.980	33.256	33.472	33.030	33.2
numba2	9.102	8.411	9.460	9.375	9.755	9.22
numba3	9.566	10.144	9.219	9.344	9.665	9.59
pypy2	8.456	7.844	8.286	8.138	7.952	8.13
pypy3	7.552	8.093	8.108	8.669	8.623	8.21
shedskin	8.024	8.070	8.399	8.126	8.277	8.18

Figure 5.21: energy consumption of different implemetations using Bit Operation benchmarks (Joule)

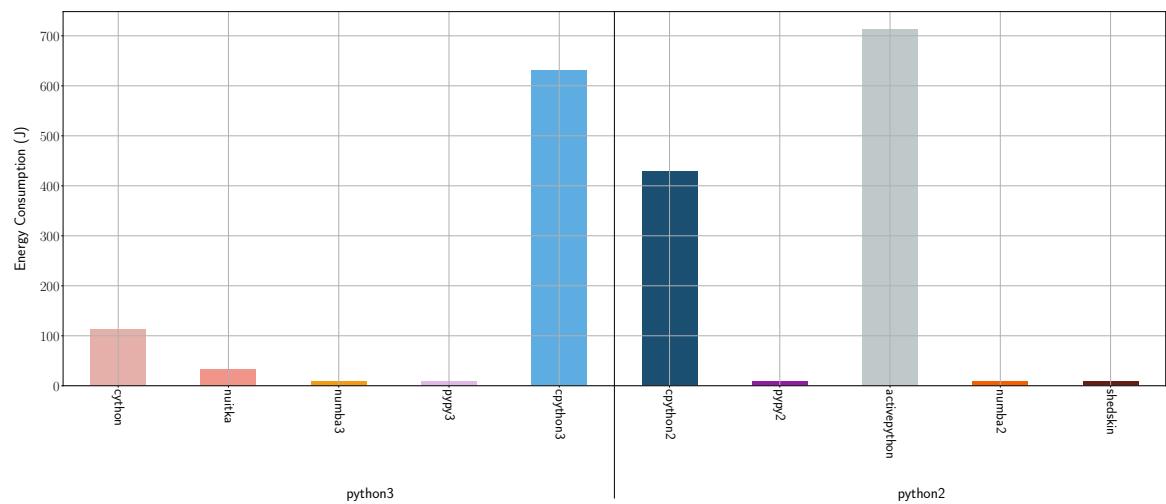


Figure 5.22: Summary of the binary operations for different python implementations

Chapter 6

The Impact of Java Virtual Machine on Energy Consumption

6.1 Introduction

As reported in the state of the art, Java is one of the most popular programming languages adopted by practitioners. In fact, in 2022, Java will be second only to Python, according to PYPL¹. Furthermore, if we take into consideration legacy applications, Java becomes the most used programming language. According to the TIOBE index, Java was the most frequently used language from 2002 until 2017, and it remained in the top 5 after that². In addition to its popularity, Java exhibits an interesting behavior when it comes to energy consumption and performance, Java applications can be at the same time one of the most energy-efficient or hungry solutions. As we have seen in the previous chapter, an inappropriate combination of parameters can drive Java applications from the top language to the bottom just by setting the wrong parameters. In this chapter, we want to dig deeper into this aspect of Java and study its runtime. This chapter thus focuses on the impact of the runtime of Java applications on energy consumption.

6.1.1 Goal

In this section, we will investigate the following research questions:

RQ 1: What is the impact of existing JVM distributions on the energy consumption of Java-based software services?

¹<https://pypl.github.io/PYPL.html>

²<https://www.tiobe.com/tiobe-index/>

RQ 2: What are the relevant JVM settings that can reduce the energy consumption of a given software service?

To answer those research questions, we conduct an empirical study to highlight the impact of this runtime. Figure

6.2 Experimental Protocol

To investigate the effect that can have the JVM distribution choice and/or parameters on software energy consumption, we conducted a wide set of experiments on a cluster of machines and used several established Java benchmarks and JVM configurations.

6.2.1 Measurement Contexts

Software Settings. For the sake of reproducibility, each experiment runs within a Docker container based on SDKMAN³ image and Alpine docker.⁴

Hardware Settings. To report on reproducible measurements, we used the cluster Dahu from the G5K platform [8] for most of our experiments. This cluster is composed of 32 identical compute nodes, which are equipped with 2 Intel Xeon Gold 6130 and 192 GB of RAM. Our experimental protocol enforces that the software under test is the only process executed on the node configured with a very minimal Linux Debian 9 (4.9.0 kernel version). The minimal OS configuration ensures that only mandatory services and daemons are kept active to conduct robust experiments and reduce the factors that can affect the energy consumption measurements during our experiments [101].

Java Virtual Machines Candidates. We considered a set of 52 JVM distributions taken from 8 different providers/packagers mostly obtained from SDKMAN, as listed in Table 6.1. Depending on providers, either all the versions, majors, or LTS are made available by SDKMAN.

6.2.2 Workload

We ran our experiments across 12 Java benchmarks we picked from OpenBenchmarking.org.⁵ This includes 5 acknowledged benchmarks from the DACAPO benchmark suite v. 9.12 [15], namely Avrora, H2, Lusearch, Sunflow and PMD, that have been widely used in previous

³<https://sdkman.io>

⁴<https://github.com/alpinelinux/docker-alpine>

⁵<https://openbenchmarking.org>

Table 6.1: List of selected JVM distributions.

Distribution	Provider	Support	Selected versions
HOTSPOT	Adopt OpenJDK	ALL	8.0.275, 11.0.9, 12.0.2, 13.0.2, 14.0.2, 15.0.1
HOTSPOT	Oracle	ALL	8.0.265, 9.0.4, 10.0.2, 11.0.2, 12.0.2, 13.0.2, 14.0.2, 15.0.1, 16.ea.24
ZULU	Azul Systems	ALL	8.0.272, 9.0.7, 10.0.2, 11.0.9, 12.0.2, 13.0.5, 14.0.2, 15.0.1
SAPMACHINE	SAP	ALL	11.0.9, 12.0.2, 13.0.2, 14.0.2, 15.0.1
LIBRCA	BellSoft	ALL	8.0.275, 11.0.9, 12.0.2, 13.0.2, 14.0.2, 15.0.1
CORRETTO	Amazon	MJR	8.0.275, 11.0.9, 15.0.1
HOTSPOT	Trava OpenJDK	LTS	8.0.232, 11.0.9
DRAGONWELL	Alibaba	LTS	8.0.272, 11.0.8
OPENJ9	Eclipse	ALL	8.0.275, 11.0.9, 12.0.2, 13.0.2, 14.0.2, 15.0.1
GRAALVM	Oracle	LTS	19.3.4.r8, 19.3.4.r11, 20.2.0.r8, 20.2.0.r11
MANDREL	Redhat	LTS	20.2.0.0

studies and proven to be accurate for memory management and computer architecture [75, 67]. It consists of open-source and real-world applications with non-trivial memory loads. Then, we also considered 7 additional benchmarks from the RENAISSANCE benchmark suite [110, 111], namely ALS, Dotty, Fj-kmeans, Neo4j, Philosophers, Reaction and Scrabble, which offer a diversified set of benchmarks aimed at testing JIT, GC, profilers, analyzers, and other tools. The benchmarks we picked from both suites exercise a broad range of programming paradigms, including concurrent, parallel, functional, and object-oriented programming. Table 6.2 summarizes the selected benchmarks with a short description. Meanwhile, Figure 6.1 highlights the scope of each benchmark from the test suite.

6.2.3 Metrics and Measurement

Since the goal of this study is the green aspect of JVM, our key metric will be the energy consumed by job completed for each JVM configuration. In addition to the energy consumption, we collected additional metrics to explain the reasons behind the behavior of each experiment. Those additional metrics are:

- *execution time,*
- *number of threads.*

Energy Measurements. We used Intel RAPL as a physical power meter to analyze the energy consumption of the CPU package and the DRAM. RAPL is one of the most accurate tools to report on the global energy consumption of a processor [68, 37]. We note that, due to CPU energy consumption variations issues [101], we used the same node for all our experiments. Moreover, we tried to be very careful, while running our experiments, not to fall in the most common benchmarking "crimes" [?]. Every single experiment, therefore, reports on energy

Table 6.2: List of selected open-source Java benchmarks taken from DACAPO and RENAISSANCE.

Benchmark	Description	Focus
ALS	Factorize a matrix using the alternating least square algorithm on spark	Data-parallel, compute-bound
Avrora	Simulates and analyses for AVR microcontrollers	Fine-grained multi-threading, events queue
Dotty	Uses the dotty Scala compiler to compile a Scala code-base	Data structure, synchronization
Fj-Kmeans	Runs K-means algorithm using a fork-join framework	Concurrent data structure, task parallel
H2	Simulates an SQL database by executing a TPC-C like benchmark written by Apache	Query processing, transactions
Lusearch	Searches keywords over a corpus of data comprising the works of Shakespeare and the King James bible	Externally multi-threaded
Neo4j	Runs analytical queries and transactions on the Neo4j database	Query Processing, Transactions
Philosophers	Solves dining philosophers problem	Atomic, guarded blocks
PMD	Analyzes a list of Java classes for a range of source code problems	Internally multi-threaded
Reactors	Runs a set of message-passing workloads based on the reactors framework	Message-passing, critical-sections
Scrabble	Solves a scrabble puzzle using Java streams	Data-parallel, memory-bound
Sunflow	Renders a classic Cornell box; a simple scene comprising two teapots and two glass spheres within an illuminated box	Compute-bound

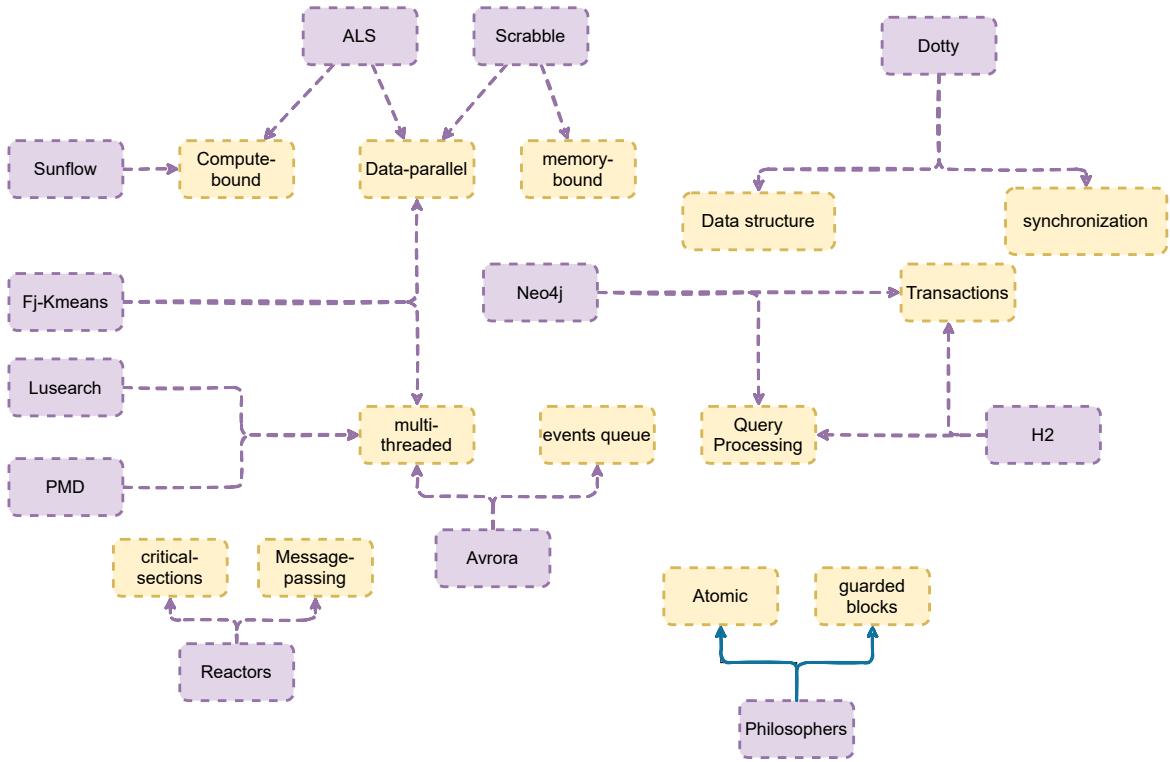


Figure 6.1: Target scope of DACAPO and RENAISSANCE benchmarks.

metrics obtained from at least 20 executions of 50 iterations per benchmark. All of our experiments are available for use/reproducibility from our anonymous repository.⁶

Number of threads. To collect the number of active threads used by the experiment, we use the command `top` and record at fixed intervals.

6.2.4 Extension

We also added an extension to the protocol to allow the user to run the same experiment with different configurations. The package is available in the GitHub repository.⁷ To add extra **jvm candidates** for the benchmark applications, we added a new configuration file `jvm.sh` in the root directory of the repository, where we put the name and the version of the jvm to be used. For the **input workload**, the benchmarks should be provided in the `benchmarks` directory. As for extra **metrics**, one can create a new script file that monitors the experiment and record the metrics. We provide some examples, such as `recordpower.sh` to measure the instantaneous power and `recordthreads.sh` to measure the number of active threads during the experiment.

⁶<https://anonymous.4open.science/r/jvm-comparaison-213E/Readme.md>

⁷<https://github.com/chakib-belgaid/jvm-comparaison>

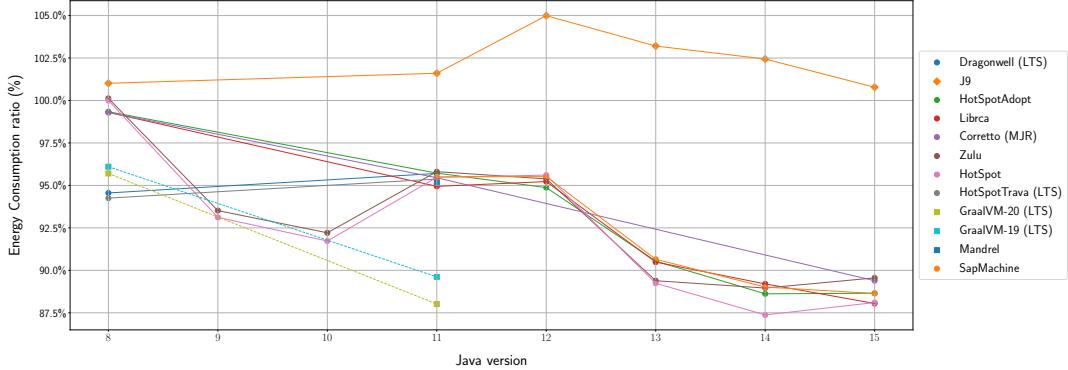


Figure 6.2: Energy consumption evolution of selected JVM distributions along versions.

For faster experiments, we propose JREFERRAL⁸, an open source tool that automatically compares the JVM energy consumption and recommends the most energy-efficient configuration for a given Java application. We further discuss this tool in Section 7.2.

6.3 Experiments & Results

6.3.1 Energy Impact of JVM Distributions

Job-oriented applications. To answer our first research question, we executed 62,400 experiments by combining the 52 JVM distributions with the 12 Java benchmarks, thus reasoning on 100 energy samples acquired for each of these combinations. Figure 6.2 first depicts the accumulated energy consumption of the 12 Java benchmarks per JVM distribution and major versions (or LTS when unavailable). Concretely, We measure the energy consumption of each of the benchmarks and compute the ratio of energy consumption compared to HOTSPOT-8, which we consider as the baseline in this experiment. Then, we sum the ratios of the 12 benchmarks and depict them as percentages in Figure 6.2.

One can observe that, along with time and versions, the energy efficiency of JVM distributions tends to improve (10% savings), thus demonstrating the benefits of optimizations delivered by the communities. Yet, one can also observe that energy consumption may differ from one distribution to another, thus showing that the choice of a JVM distribution may have a substantial impact on the energy consumption of the deployed software services. For example, one can note that J9 can exhibit up to 15% of energy consumption overhead, while other distributions seem to converge towards a lower energy footprint for the latest version

⁸<https://github.com/chakib-belgaid/jreferral>

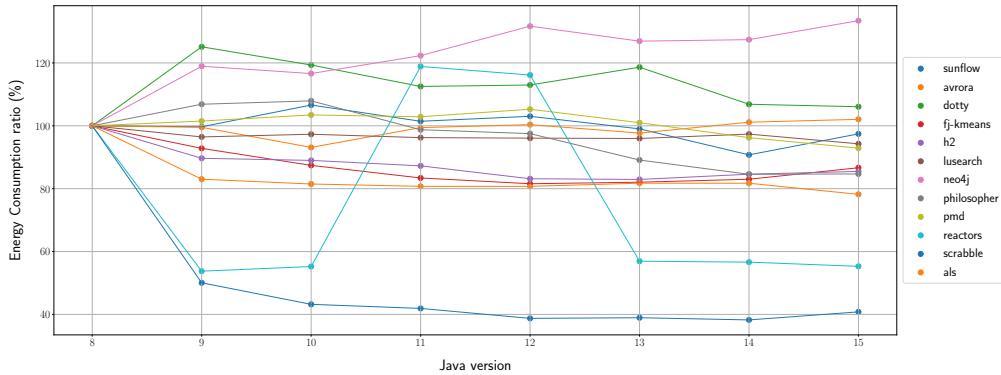


Figure 6.3: Energy consumption of the HotSpot JVM along versions.

of Java. As GRAALVM adopts a different strategy focused on LTS support, one can observe that its recent releases provide the best energy efficiency for Java 11, but recent releases of other distributions seem to reach similar efficiency for Java 13 and above, which are recent versions not supported by GRAALVM yet.

Interestingly, this convergence of distributions has been observed since Java 11 and coincides with the adoption of DCE VM by HOTSPOT. Ultimately, 3 clusters of JVMs that encompass JVMs with similar energy consumption can be seen through Figure 6.2: J9, the HOTSPOT and its variants, and GRAALVM. Additional detailed figures to illustrate the evolution of energy consumption per benchmark/JVM are made available from the online repository.⁹

Then, Figure 6.3 depicts the evolution of the energy consumption of the 12 benchmarks, when executed on the HOTSPOT JVM. Figure 6.3 reports on the energy consumption variation of individual benchmarks, using HOTSPOT-8 as the baseline. Our results show that the JVM version can severely impact the energy consumption of the application. However, unlike Figure 6.2, one can observe that, depending on applications, the latest JVM versions can consume less energy (60% less energy for Scrabble) or more energy (25% more energy for the Neo4J). It is worth noticing that the energy consumption of some benchmarks, such as Reactors, exhibit large variations across JVM versions due to experimental features and changes that are not always kept when releasing LTS versions (version 11 here). For example, the introduction of VarHandle to allow low-level access to the memory order modes available in JDK 9¹⁰ and work along Unsafe Class was removed from JVM 11.¹¹

⁹<https://github.com/chakib-belgaid/jvm-comparaison>

¹⁰<https://gee.cs.oswego.edu/dl/html/j9mm.html>

¹¹<https://blogs.oracle.com/javamagazine/the-unsafe-class-unsafe-at-any-speed>

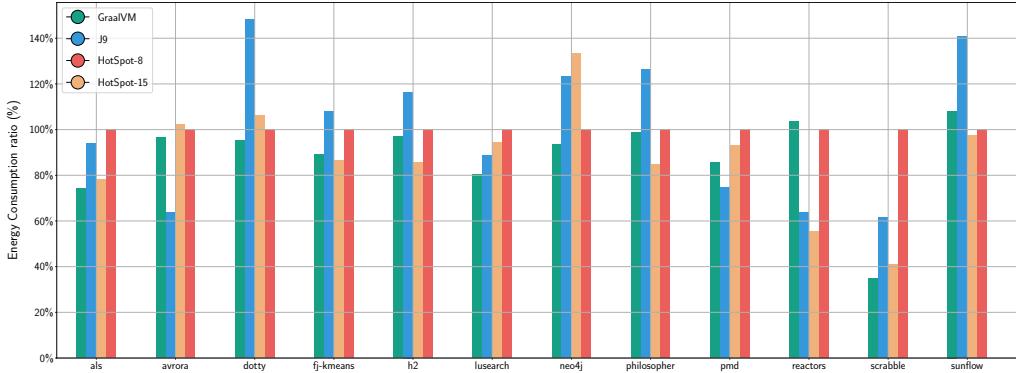


Figure 6.4: Energy consumption comparison across Java benchmarks for HOTSPOT, GRAALVM & J9.

Given that the wide set of distributions and versions seems to highlight 3 classes of energy behaviors, the remainder of this chapter considers the following distributions as relevant samples of JVM to be further evaluated: 20.2.0.r11-grl (GRAALVM), 15.0.1-open (HOTSPOT-15), 15.0.21.j9 (J9). We also decided to keep the 8.0.275-open (HOTSPOT-8) as a baseline JVM for some figures to highlight the evolution of energy consumption over time/versions.

Figure 6.4 further explores the comparison of energy efficiency of the JVM distributions per benchmark. One can observe that, depending on the benchmark's focus, the energy efficiency of JVM distributions may strongly vary. When considering individual benchmarks, J9 performs the worst for at least 6 out of 12 benchmarks—i.e., the worst ratio among the 4 tested distributions. Even though, J9 can still exhibit a significant energy saving for some benchmarks, such as Avrora, where it consumes 38% less energy than HOTSPOT and others.

Interestingly, GRAALVM delivers good results overall, being among the distributions with low energy consumption for all benchmarks, except for Reactors and Avrora. Yet, some differences still can be observed with HOTSPOT depending on applications. The newer version of HOTSPOT-15 was averagely good and, compared to HOTSPOT-8, it significantly enhances energy consumption for most scenarios. Finally, Neo4J is the only selected benchmark where HOTSPOT-8 is more energy efficient than HOTSPOT-15.

Service-oriented applications. In this section, instead of considering bounded execution of benchmarks, we run the same benchmarks as services for 20 minutes, and we compare the average power and total requests processed by each of the 3 JVM distributions. Globally, the results showed that the average power when using GRAALVM, HOTSPOT, and OPENJ9 is often equivalent and stable over time. This means that the energy efficiency observed for some JVM distributions with job-oriented applications is mainly related to shorter execution times, which incidentally results in energy savings. Nonetheless, we can highlight two interesting

Table 6.3: Power per request for HOTSPOT, GRAALVM & J9.

Benchmark	JVM	Power (P)	Requests (R)	$P/R \times 10^{-3}$
Scrabble	GRAALVM	109 W	5,336 req	20 mW
	HOTSPOT	98 W	3,595 req	27 mW
	J9	92 W	2,603 req	35 mW
Dotty	GRAALVM	45 W	510 req	88 mW
	HOTSPOT	45 W	597 req	75 mW
	J9	46 W	381 req	120 mW

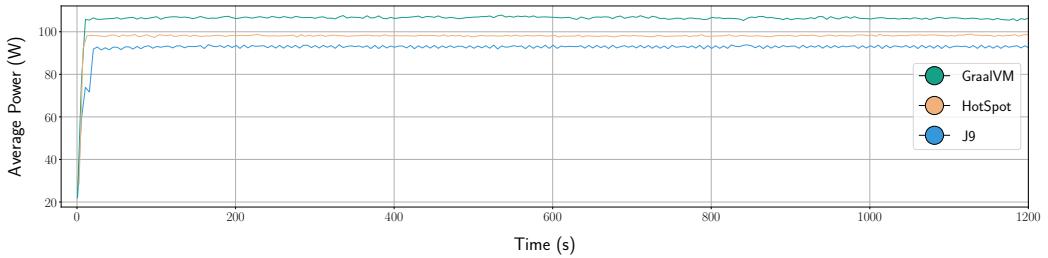


Figure 6.5: Power consumption of Scrabble as a service for HOTSPOT, GRAALVM & J9.

observations for two benchmarks whose behaviors differ from others. First, the analysis of the Scrabble benchmark experiments showed that, in some scenarios, some JVMs can exhibit different power consumptions. Figure 6.5 depicts the power consumed by the 3 JVM distributions for the Scrabble benchmark. One can clearly see that GRAALVM requires an average power of 109 W, which is 9 W higher than HOTSPOT-15 and 15 W higher than J9. When it comes to the number of requests processed by Scrabbles during that same amount of time, GRAALVM completes 5,336 requests, against 3,595 for HOTSPOT and 2,603 for J9, as shown in Table 6.3. The higher power usage for GRAALVM helped in achieving a high amount of requests, but also the fastest execution of every request, which was 40% faster on GRAALVM. Thus, GRAALVM was more energy efficient, even if it uses more power, which confirms the results observed in Figure 6.4 for this benchmark.

The second interesting situation was observed on the Dotty benchmark. More specifically, during the first 100 seconds of the execution of the Dotty benchmark on all evaluated JVMs. At the beginning of the execution, GRAALVM has a slightly lower power consumption, is faster, and consumes 10% less energy. After about 150 seconds, the power differences between the 3 JVMs is barely noticeable. One can, however, notice the effect of the JIT, as HOTSPOT takes the advantage over GRAALVM and becomes more energy efficient. In total, HOTSPOT completes 597 requests against 510 for GRAALVM and 381 for J9, as shown in Table 6.3. HOTSPOT was thus the best choice in the long term, which explains why it is always necessary to consider a warm-up phase and wait for the JIT to be triggered before

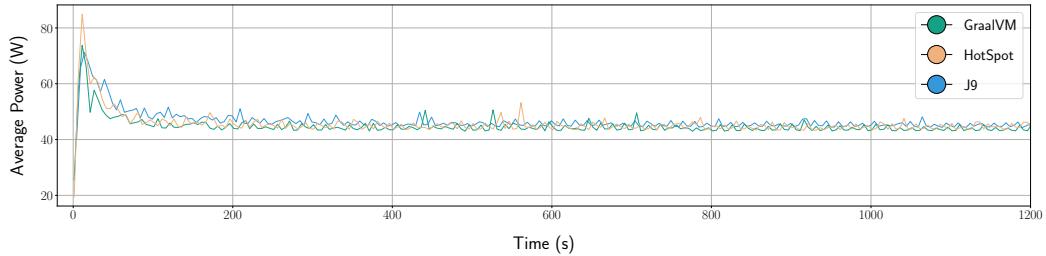


Figure 6.6: Power consumption of Dotty as a service for HOTSPOT, GRAALVM & J9.

evaluating the effect of the JVM or the performance of an application. This is precisely what we did in our experiments, and it is why HOTSPOT was more energy efficient than GRAALVM in Figure 6.4; therefore, ignoring the warm-up phase would have been misleading.

To answer RQ 1, we conclude that—while most of the JVM platforms perform similarly—we can cluster JVMs in 3 classes: HOTSPOT, J9, and GRAALVM. The choice of one JVM of these classes can have a major impact on software energy consumption, which strongly depends on the application context. When it comes to the JVM version, the latest releases tend to offer the lowest power consumption, but experimental features should be carefully configured, thus further questioning the impact of JVM parameters.

6.3.2 Energy Impact of JVM Settings

The purpose of our study is not only to investigate the impact of the JVM platform on energy consumption, but also the different JVM parameters and configurations that might have a positive or negative effect, with a focus on 3 available settings: multi-threading, JIT, and GC.

Multithreading

The purpose of this phase is to investigate the impact JVM thread management strategies on energy consumption. This encompasses exploring if the management strategies of application-level parallelism (so-called threads) result in different energy efficiencies, depending on JVM distributions.

Investigating such a hypothesis requires a selection of highly parallel and CPU-intensive benchmarks, which is one of the main criteria for our benchmark selection. As no tool can accurately monitor the energy consumption at a thread level, we monitor the global power consumption and CPU utilization during the execution using RAPL for the energy, and several Linux tools for the CPU-utilization (`htop`, `cpufreq`). Knowing that most of the benchmarks are multi-threaded jobs that use multiple cores, further analysis of thread



Figure 6.7: Active threads evolution when using HOTSPOT, GRAALVM, or J9.

management is required to understand the results of our previous experiments. We thus selected the benchmarks that highlighted the highest differences along JVM distributions from Figure 6.4, namely Avrora and Reactors. We studied their multi-threaded behavior to optimize their energy efficiency.

Figure 6.7 delivers a closer look to the thread allocation strategies adopted by JVM. First, Figure 6.7a illustrates the active threads count evolution over time (excluding the JVM-related threads, usually 1 or 2 extra threads depending on the execution phase) for Avrora. One can notice through the figure that J9 exploits the CPU more intensively by running much more parallel threads compared to other JVMs (an average of 5.1 threads per second for J9 while the other JVMs do not exceed 1.5 thread per second). Furthermore, the number of context switches is twice bigger for J9, while the number of soft page faults is twice smaller. The efficient J9 thread management explains why running the Avrora benchmark took much less time and consumed less energy, given that no other difference for the JIT or GC configuration was spotted between the JVMs. Another key reason for the J9's efficiency

for the Aurora benchmark is memory allocation, as OpenJ9 adopts a different policy for the heap allocation. It creates a non-collectible thread local heap (TLH) within the main heap for each active thread. The benefit of cloning a dedicated TLH is the fast memory access for independent threads: each thread has its heap and no deadlock can occur.

The second example in Figure 6.7b depicts the active threads evolution over time of the Reactors benchmark. In this case, all the JVMs have a close average of threads per second. Nevertheless, one can still observe that HOTSPOT-15 and J9 keep running faster, which confirms the results of Figure 6.4, where both JVMs consume much less energy compared to GRAALVM and HOTSPOT-8. This difference in energy consumption between benchmarks can be less likely caused by thread management for the Reactors benchmark, as HOTSPOT-8 reports on a higher average of active threads. However, the TLH mechanism was not as efficient as for the Aurora benchmark, as dedicating a heap for each thread can also cause some extra memory usage for data duplication and synchronization, especially if a lot of data is shared between threads.

In conclusion, JVMs thread management can sometimes constitute a key factor that impacts software energy consumption. However, we suggest checking and comparing JVMs before deploying a software, especially if the target application is parallel and multi-threaded.

Just-in-Time Compilation

The purpose of experiments on JIT is to highlight the different strategies that can impact software energy consumption within a JVM and between JVMs. We identified a set of JIT compiler parameters for every JVM platform.

For J9, we considered fixing the intensity of the JIT compiler at multiple levels (cold, warm, hot, veryhot, and scorching).¹² The hotter the JIT, the more code optimization to be triggered. We also varied the minimum count method calls before a JIT compilation occurs (10, 50, 100), and the number of JIT instances threads (from 1 to 7). For HOTSPOT-15, we conducted experiments while disabling the tiered compilation (that generates compiled versions of methods that collect profiling information about themselves), and we also varied the JIT maximum compilation level from 0 to 4, we also tried out HOTSPOT with a basic GRAALVM JIT. We note that level 0 of JIT compilation only uses the interpreter, with no real JIT compilation. Levels 1, 2, and 3 use the C1 compiler (called client-side) with different amounts of extra tuning. The JIT C2 (also called server-side JIT) compiler only kicks in at level 4.

For GRAALVM, we conducted experiments with and without the JVMBI (a Java-based JVM compiler interface enabling a compiler written in Java to be used by the JVM as a

¹²[<https://www.eclipse.org/openj9/docs/jit/>]

dynamic compiler). We also considered both the community and economy configurations (no enterprise). A JIT+AOT (Ahead Of Time) disabling experiment has also been considered for all of the 3 JVM platforms. Table 6.4 reports on the energy consumption of the experiments we conducted for most of the benchmarks and JIT configurations under study.

The p-values are computed with the Mann-Whitney test, with a null hypothesis of the energy consumption being equal to the default configuration. The p-values in bold show the values that are significantly different from the default configuration with a 95% confidence, where the values in green highlight the strategies that consumed significantly less energy than the default (less energy and significant p-value).

For J9, we noticed that adopting the default JIT configuration is always better than specifying a custom JIT intensity. The warm configuration delivers the closest results to the best results observed with the default configuration. Moreover, choosing a low minimum count of method calls seems to have a negative effect on the execution time and energy consumption. The only parameter that can give better performance than the default configuration in some cases is the number of parallel JIT threads—using 3 and 7 parallel threads—but is not statistically significant.

For GRAALVM, the default community configuration is often the one that consumes the least energy. Disabling the JVMCI can—in some cases—have a benefit (16% of energy consumption reduction for the H2 benchmark), but still gave overall worst results (80% more energy consumption for the Neo4J benchmark). In addition, switching the economy version of the GRAALVM JIT often results in consuming more energy and delaying the execution.

For HOTSPOT, keeping the default configuration of the JIT is also mostly good. The usage of the C2 JIT is often beneficial (JIT level 4) in most cases while using the GRAALVM JIT reported similar energy efficiency. Yet, some benchmarks showed that using only the C1 JIT (JIT level 1) is more efficient and even outperforms the usage of the C2 compiler. 10% on Avrora and 30% on Pmd are examples of energy savings observed by using the C1 compiler. However, being limited to the C1 compiler can also cause a huge degradation in energy consumption, such as 32% and 34% of additional energy consumed for the Dotty and FJ-kmeans benchmarks, respectively. Hence, if it is a matter of not using the C2 JIT, the experiments have shown that the level 1 JIT is always the best, compared to levels 2 or 3 that also use the C1 JIT, but with more options, such as code profiling that impacts negatively the performance and the energy efficiency. Level 0 JIT compilation should never be an option to consider. No p-value has been computed for Level 0, due to the limited amount of iterations executed with this mode (very high execution time, clearly much more consumed energy).

Globally, we conclude through these experiments that keeping the default JIT configuration was more energy efficient in 80% of our experiments and for the 3 classes of JVMs.

Table 6.4: Energy consumption when tuning JIT settings on HOTSPOT, GRAALVM & J9

JVM	Mode	ALS		Avrora		Dotty		Fj-kmeans		H2	
GRAALVM	<i>Default</i>	2848	<i>p-values</i>	3861	<i>p-values</i>	2271	<i>p-values</i>	948	<i>p-values</i>	1959	<i>p-values</i>
	DisableJVMCI	3099	0.001	4012	0.041	2694	0.001	934	0.011	1771	0.005
	Economy	4503	0.001	3895	0.793	3466	0.001	1306	0.002	2560	0.001
J9	<i>Default</i>	3792	<i>p-values</i>	2122	<i>p-values</i>	3515	<i>p-values</i>	1271	<i>p-values</i>	2426	<i>p-values</i>
	Thread 1	4157	0.001	2121	0.875	4749	0.001	1297	0.097	2597	0.066
	Thread 3	3849	0.018	2105	0.713	3574	0.104	1259	0.371	2450	0.637
	Thread 7	3843	0.041	2386	0.372	3511	0.875	1259	0.25	2424	0.637
	Count 0	8461	0.001	2425	0.001	4877	0.001	2289	0.002	3212	0.001
	Count 1	4281	0.001	2150	0.431	3164	0.001	1841	0.002	2546	0.431
	Count 10	3980	0.001	2431	0.713	3771	0.001	1312	0.011	2779	0.003
	Count 100	3878	0.007	2141	0.713	3469	0.227	1363	0.523	2513	0.128
	Cold	6788	0.001	2134	0.637	4855	0.001	1636	0.002	2873	0.001
	Warm	4594	0.001	2112	0.713	4253	0.001	1244	0.055	2521	0.128
	Hot	7553	0.001	2310	0.001	12749	0.001	1452	0.002	3973	0.001
	VeryHot	15113	0.001	3300	0.001	18235	0.001	2430	0.002	7205	0.001
	Schorching	18316	0.001	3541	0.001	21686	0.001	2514	0.002	7855	0.001
HOTSPOT	<i>Default</i>	2997	<i>p-values</i>	4014	<i>p-values</i>	2516	<i>p-values</i>	934	<i>p-values</i>	1796	<i>p-values</i>
	Graal	2999	0.637	3971	0.318	2512	0.318	929	0.609	1662	0.007
	Lvl 0	491443	/	14484	/	84395	/	/	/	52344	/
	Lvl 1	/	/	3731	0.001	3302	0.001	1256	0.002	2523	0.001
	Lvl 2	3079	0.004	4110	0.189	3723	0.001	22547	0.002	2840	0.001
	Lvl 3	16375	0.001	7729	0.001	6789	0.001	144914	0.002	4139	0.001
	NotTired	3254	0.001	3901	0.189	3110	0.001	912	0.021	1846	0.227
JVM	Mode	Neo4j		Pmd		Reactors		Scrabble		Sunflow	
GRAALVM	<i>Default</i>	3313	<i>p-values</i>	297	<i>p-values</i>	23452	<i>p-values</i>	452	<i>p-values</i>	335	<i>p-values</i>
	DisableJVMCI	5086	0.001	353	0.001	25007	0.007	503	0.002	354	0.227
	Economy	9525	0.001	270	0.001	30317	0.001	649	0.002	392	0.002
J9	<i>Default</i>	4336	<i>p-values</i>	277	<i>p-values</i>	12705	<i>p-values</i>	734	<i>p-values</i>	476	<i>p-values</i>
	Thread 1	4906	0.001	350	0.001	12800	0.713	948	0.002	626	0.005
	Thread 3	4477	0.005	294	0.004	12647	0.875	795	0.021	457	0.27
	Thread 7	4431	0.104	273	0.372	12600	0.875	808	0.055	463	0.372
	Count 0	10565	0.001	744	0.001	18084	0.001	1476	0.002	922	0.001
	Count 1	7166	0.001	272	0.128	14715	0.001	1005	0.002	514	0.052
	Count 10	4979	0.001	299	0.001	12000	0.104	860	0.005	1182	0.001
	Count 100	4547	0.001	262	0.031	12313	0.024	768	0.16	634	0.004
	Cold	7250	0.001	275	0.372	20380	0.001	870	0.005	386	0.001
	Warm	5305	0.001	411	0.001	13726	0.001	913	0.002	336	0.001
	Hot	8979	0.001	857	0.001	36534	0.001	1180	0.002	506	0.128
	VeryHot	19359	0.001	793	0.001	38303	0.001	5420	0.002	1692	0.001
	Schorching	26409	0.014	808	0.001	43929	0.001	5583	0.002	1778	0.001
HOTSPOT	<i>Default</i>	4787	<i>p-values</i>	323	<i>p-values</i>	11685	<i>p-values</i>	530	<i>p-values</i>	325	<i>p-values</i>
	Graal	4750	0.372	327	0.189	11548	0.523	537	0.701	338	0.564
	Lvl 0	356287	/	1073	/	148381	/	/	/	14559	/
	Lvl 1	8304	0.001	222	0.001	22410	0.002	735	0.002	277	0.007
	Lvl 2	19058	0.001	226	0.001	40701	0.002	2291	0.002	4131	0.001
	Lvl 3	44594	0.001	330	0.005	190124	0.002	9070	0.002	10449	0.001
	NotTired	3844	0.001	933	0.001	11256	0.041	588	0.003	405	0.001

Table 6.5: The different J9 GC policies

Policy	Description
Balanced	Evens out pause times & reduces the overhead of the costlier operations associated with GC
Metronome	GC occurs in small interruptible steps to avoid stop-the-world pauses
Nogc	Handles only memory allocation & heap expansion, with no memory reclaim
Gencon (default)	Minimizes GC pause times without compromising throughput, best for short-lived objects
Concurrent Scavenge	Minimizes the time spent in stop-the-world pauses by collecting nursery garbage in parallel with running application threads
optthruput	Optimized for throughput, stopping applications for long pauses while GC takes place
Optavgpause	Sacrifices performance throughput to reduce pause times compared to optthruput

This advocates using the default JIT configuration that can often deliver near-optimal energy efficiency. Although, some other configurations, such as using only the C1 JIT or disabling the JVMCI could be advantageous in some cases.

Garbage Collection

Changing or tuning the GC strategy has been acknowledged to impact the JVM performances [77]. To investigate if this impact also benefits energy consumption, we conducted a set of experiments on the selected JVMs. We considered different garbage collector strategies with a limited memory quantity of 2 GB, and recorded the execution time and the energy consumption. The tested GC strategies options mainly vary between J9 and the other 2 JVMs, as detailed in Table 6.5.

For HOTSPOT and GRAALVM, we also considered many GC policies, as described in Table 6.6. Furthermore, other GC settings have also been tested for all JVM platforms, such as the pause time, the number of parallel threads and concurrent threads and tenure age.

Table 6.7 summarizes the results of all the tested GC strategies with our selected benchmarks and the p-values of the Mann-Whitney test, with a null hypothesis of the energy consumption being equal to the default configuration with a 95% confidence. The p-values in bold show the values that are significantly different from the default configuration, whereas the values in green highlight the strategies that consumed significantly less energy than the default. For GRAALVM, one can see that the GC default configuration is efficient in most experiments, compared to other strategies. The main noticeable impact is related to the ParallelGC and ParallelOldGC. In fact, the ParallelGC can be 13% more energy efficient in

Table 6.6: The different HOTSPOT/GRAALVM GC policies

Policy	Description
G1GC (default)	Uses concurrent & parallel phases to achieve low-pauses GC and maintain good throughput
SerialGC	Uses a single thread to perform all garbage collection work (no threads communication overhead)
ParallelGC	Known as throughput collector: similar to SerialGC, but uses multiple threads to speed up garbage collections for scavenges
parallelOldGC	Use parallel garbage collection for the full collections, enabling it automatically enables the ParallelGC

some applications with a significant p-value, such as Reactors, compared to default. However, the same GC strategy can cause the software to consume twice times more, as for the Neo4j benchmark, due to the high communications between the GC threads, and the fragmentation of the memory.

For J9, the default Gencon GC causes the software to report an overall good energy efficiency among the tested benchmarks. However, other GC can cause better or worse energy consumption than Gencon depending on workloads. Using the Metronome GC consumes 35% less energy for the ALS benchmark and 17% less energy for the Sunflow benchmark, but it also consumes 100% more energy for the Neo4j benchmark and 28% more energy for Reactors. The reason is that Metronome occurs in small preemptible steps to reduce the GC cycles composed of many GC quanta. This suits well for real-time applications and can be very beneficial when long GC pauses are not desired, as observed for ALS. However, if the heap space is insufficient after a GC cycle, another cycle will be triggered with the same ID. As Metronome supports class unloading in the standard way, there might be pause time outliers during GC activities, inducing a negative impact on the Neo4j execution time and energy consumption.

The same goes for the Balanced GC that tries to reduce the maximum pause time on the heap by dividing it into individually managed regions. The Balanced strategy is preferred to reduce the pause times that are caused by global GC, but can also be disadvantageous due to the separate management of the heap regions, such as for ALS where it consumed about three times the energy consumption, compared to the default Gencon GC. On the other hand, the Optthruput GC, which stops the application longer and less frequently, gave very good overall results and sometimes even outperformed the Gencon GC by a small margin. Other JVM parameters, such as the ConcurrentScavenge or noAdaptiveTenure did not have a substantial impact during our experiments.

Table 6.7: Energy consumption when tuning GC settings on HOTSPOT, GRAALVM & J9

JVM	Mode	ALS		Avrora		Dotty		H2		Neo4j	
GRAALVM	<i>Default</i>	2570	<i>p-values</i>	4153	<i>p-values</i>	2223	<i>p-values</i>	1870	<i>p-values</i>	5256	<i>p-values</i>
	1Concurrent	2567	0.403	4007	0.023	2220	1.000	1883	0.982	5368	1.000
	1Parallel	2668	0.012	3904	0.008	2228	0.835	2022	0.000	5836	0.012
	5Concurrent	2570	0.676	4117	0.161	2215	0.210	1862	0.505	5259	1.000
	5Parallel	2561	0.676	3863	0.012	2237	1.000	1910	0.103	5223	0.403
	DisableExplicitGC	2559	0.210	3911	0.003	2215	1.000	1978	0.018	5106	0.210
	ParallelICG	2720	0.012	4016	0.206	2237	0.531	1945	0.000	13172	0.037
	ParallelOldGC	2715	0.012	4032	0.103	2221	1.000	1925	0.002	13362	/
	<i>Default</i>	3371	<i>p-values</i>	2243	<i>p-values</i>	3237	<i>p-values</i>	2107	<i>p-values</i>	6277	<i>p-values</i>
J9	Balanced	9012	0.012	2232	0.597	3429	0.012	2247	0.002	8853	0.012
	ConcurrentScavenge	3487	0.012	2270	0.280	3388	0.012	2319	0.001	6857	0.012
	Metronome	2098	0.012	2265	0.505	3815	0.012	2717	0.000	12103	0.012
	Nogc	3454	0.022	2239	0.872	3259	0.144	2207	0.031	61781	0.012
	Optavgpause	3601	0.012	2431	0.370	3425	0.012	2169	0.297	7495	0.012
	Optthruput	3357	1.000	2432	0.241	3178	0.403	2194	0.139	6324	0.835
	ScvNoAdaptiveTenure	3494	0.012	2253	0.800	3248	0.835	2161	0.103	8442	0.012
	<i>Default</i>	2765	<i>p-values</i>	4115	<i>p-values</i>	2492	<i>p-values</i>	1673	<i>p-values</i>	8152	<i>p-values</i>
	1Concurrent	2775	0.060	4137	0.346	2493	0.676	1675	0.918	8062	0.531
HOTSPOT	1Parallel	2863	0.012	4142	0.800	2526	0.037	1853	0.001	8270	0.676
	5Concurrent	2758	0.676	4091	0.872	2485	0.296	1681	0.608	8087	0.835
	5Parallel	2767	0.144	4176	0.077	2473	0.060	1654	0.720	8046	0.835
	DisableExplicitGC	2734	0.012	4062	0.448	2483	0.835	1702	0.248	7710	0.037
	ParallelICG	2653	0.012	4064	0.629	2356	0.012	1602	0.008	8953	0.060
	ParallelOldGC	2764	0.531	4070	0.872	2525	0.802	1675	0.959	7963	0.403
	SerialGC	2593	0.012	4083	0.395	2378	0.012	1620	0.046	5745	0.012
	<i>Default</i>	232	<i>p-values</i>	1644	<i>p-values</i>	589	<i>p-values</i>	510	<i>p-values</i>	510	<i>p-values</i>
	Balanced	235	0.412	1902	0.020	661	0.061	519	0.505	519	0.505
J9	ConcurrentScavenge	233	0.878	1705	0.903	639	0.194	546	0.018	546	0.018
	Metronome	239	0.022	2089	0.020	758	0.030	422	0.000	422	0.000
	Nogc	227	0.151	1505	0.066	711	0.030	499	0.720	499	0.720
	Optavgpause	253	0.000	1772	0.391	1089	0.030	478	0.046	478	0.046
	Optthruput	232	0.878	1554	0.111	640	0.194	429	0.000	429	0.000
	ScvNoAdaptiveTenure	228	0.137	1908	0.020	618	0.665	528	0.218	528	0.218
	<i>Default</i>	316	<i>p-values</i>	1546	<i>p-values</i>	484	<i>p-values</i>	347	<i>p-values</i>	347	<i>p-values</i>
	1Concurrent	316	0.383	1533	0.665	478	0.470	334	0.218	334	0.218
	1Parallel	334	0.000	1747	0.030	592	0.030	320	0.002	320	0.002
HOTSPOT	5Concurrent	314	0.330	1497	0.665	469	0.030	336	0.259	336	0.259
	5Parallel	316	0.573	1546	0.470	489	0.470	342	0.573	342	0.573
	DisableExplicitGC	312	0.200	1545	0.470	470	0.061	325	0.014	325	0.014
	ParallelICG	300	0.000	1476	0.885	579	0.030	336	0.081	336	0.081
	ParallelOldGC	314	0.720	1582	0.194	475	0.470	333	0.151	333	0.151
	SerialGC	307	0.002	1672	0.061	601	0.030	352	0.473	352	0.473

Finally, the results of HOTSPOT shared similarities with GRAALVM. The ParallelGC happened to give better (6% for Dotty) or worst (10% for Neo4j) energy efficiency compared to the default GC. On the other hand, ParallelOldGC and Serial GC gave better results than the default G1 GC. More specifically, the second one consumed 30% and 6% less energy than the default GC for the Neo4j and Dotty benchmarks, respectively. The most interesting result for HOTSPOT is the 30% energy reduction obtained with the Serial GC. This last was also more efficient on ALS (6% less energy), compared to the default G1 GC, due to its single-threaded GC that only uses one CPU core.

Unfortunately, we cannot convey predictive patterns on how to configure the GC to optimize energy efficiency. However, some considerations should be taken into account when choosing the GC, such as the garbage collection time, the throughput, etc. Other settings are less trivial to determine, such as tenure age, memory size, and GC threads count. Experiments should thus be conducted on the software to tune the most convenient GC configuration to achieve better energy efficiency in production.

Therefore, we noticed during our experiments that, even if using the default GC configuration ensures an overall steady and correct energy consumption, we still found other settings that reduce that energy consumption in 50% of our experiments. Tuning the GC according to the hosted app/benchmark is thus critical to reducing energy consumption.

To answer RQ 2, we conclude that users should be careful while choosing and configuring the garbage collector as substantial energy enhancements can be recorded from one configuration to another. The default GC consumes more energy than other strategies in most situations. However, keeping the default JIT parameters often delivers near-optimal energy efficiency. In addition, the JVM platforms can handle differently multi-threaded applications and thus consume a different amount of time/energy. Dedicated performance tuning evaluations should therefore be conducted on such software to identify the most energy-efficient platform and settings.

6.4 Threats to Validity

Several issues may affect the validity of our work. First, we have the use of the Intel RAPL, one of the most accurate available tools to measure the energy consumption of software [68, 37]. However, RAPL only gives the global energy consumption and no fine-grained measures at process or thread levels. We used bare-metal hardware with a minimal OS and turned off all the non-essential services and daemons to limit the overhead that the OS may add to the execution, even if it is not substantial [101].

Another measurement issue is the CPU energy variation within machines (cf. Chapter 3), thus we executed all the comparable tests on the same node and with the recommended settings to mitigate this threat.

Benchmarks' execution time could also constitute a more subtle threat to the validity of our work, especially for some benchmarks that run fast, such as the Pmd benchmark. We thus gave a lot of attention to how long the benchmark is running for the hardware we used, and we tuned the input data workloads to execute benchmarks for at least many (from 10 to hundreds) seconds. Experiments ran at least 30 times to compute the average consumption and the associated standard deviation, therefore reasoning over reasonable dispersion around the average.

How generalizable are our results? We believe that our study conclusions and guideline remain empirical, as we do not intend to generalize any result we obtained for some JVM or benchmark. We provide practitioners with some prerequisites to check before software deployment to reduce the software energy footprint by considering the JVM and its settings.

6.5 Tools and contributions

J-Referral is an open-source tool designed to assist developers and practitioners in selecting the most energy-efficient JVM configuration for their software.. Table 6.8 illustrates an example of the final report returned by J-Referral. The tool was tested for 2 Java projects: Zip4J¹³ and K-nucleotide.¹⁴ Zip4J runs a large file compression, while K-nucleotide extracts a DNA sequence, and updates a hashtable of k-nucleotide keys to count specific values. The short report presented in Table 6.8 shows the ratio of potential energy saving between the most and least energy consuming tested JVM (40% and 70% energy savings for Zip4J and K-nucleotide, respectively). Options are available for J-Referral to obtain much more detailed reports including execution time, DRAM usage, split DRAM vs. CPU consumption, etc. The tool is available as open-source software (OSS) from our GitHub repository.¹⁵

6.6 Conclusion

The results of our investigations show that many JVMs share energy efficiencies and can be grouped into 3 classes: HOTSPOT, J9, and GRAALVM. The 3 selected JVM classes

¹³<https://github.com/srikanth-lingala/zip4j>

¹⁴<https://benchmarksgame-team.pages.debian.net/benchmarksgame/performance/knucleotide.html>

¹⁵<https://github.com/chakib-belgaid/jreferral>

Table 6.8: J-Referral recommendations.

Project	Metric	Energy	JVM	Execution flags
Zip4J	Least energy	2210 J	16-sapmchn	default
	Most energy	3680 J	8.0.292-J9	default
K-nucl	Least energy	1296 J	21.1.r16-grl	default
	Most energy	4433 J	15.0.1-J9	-Xjit:optlevel=cold

can however report a different energy efficiency for different software and/or workloads, sometimes by a large margin. While we did not observe a unique champion when it comes to energy consumption, GRAALVM reported the best energy efficiency for a majority of benchmarks. Nonetheless, each JVM can achieve the best or the worst depending on the hosted application. One cause can be thread management strategies, as observed with J9 when advantageously running Avrora. Moreover, some JVM settings can cause energy consumption variations. Our experiments showed that the default JIT compiler of the JVM is often near-optimal, in at least 80% of our experiments. The default GC, however, was outperforming alternative strategies in half of our experiments, with some large gains observed when using some alternative GC depending on the application characteristics.

Our main conclusions and guidelines can be thus summarized as: i) testing software on the 3 classes of JVM and identifying the one that consumes the least is a good practice, especially for multi-threading purposes, ii) while the JVM default JIT give often good energy consumption results, some settings may improve the energy consumption and could be tested, iii) the choice of the GC may lead to a large impact on the energy consumption in many situations, thus encouraging a careful tuning of this parameter before deployment. To ease the integration of the above guidelines, we propose a tool, named J-Referral, to recommend the most energy-efficient JVM distribution and configuration among more than a hundred considered possibilities. It establishes a full report on the energy consumption of both CPU and DRAM components for each JVM distribution and/or configuration to help the user to choose the one with the least consumption for Java software.

Chapter 7

Discussion and Conclusion

7.1 Conclusion

7.2 Summary of Contributions

This section will describe the contributions of this thesis. These can be summarized as follows:

1. ***First Idea:*** We proposed ...
2. ***Second Idea:*** We investigated ...
3. ***Third Idea:*** We addressed ...

7.3 Limitations and Challenges

7.4 Future Work

.... Some potential areas for future efforts could include the following:

1. ...
2. ...
3. ...

Bibliography

- [1] (2008). Pearson's Correlation Coefficient. *Springer Netherlands*.
- [2] Abuabdo, A. and Al-Sharif, Z. A. (2019). *Virtualization vs. Containerization: Towards a Multithreaded Performance Evaluation Approach*. In 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), pages 1–6.
- [3] Acun, B., Miller, P., and Kale, L. V. (2016). *Variation Among Processors Under Turbo Boost in HPC Systems*.
- [4] Akeret, J., Gamper, L., Amara, A., and Refregier, A. (2015). HOPE: A Python just-in-time compiler for astrophysical computations. *Astronomy and Computing*, 10:1–8.
- [Bailey et al.] Bailey, D., Barszcz, E., Barton, J., Browning, D., Carter, R., Dagum, L., Fineberg, S., Frederickson, P., Lasinski, T., Schreiber, R., Simon, H., Venkatakrishnan, V., and Weeratunga, S. *The NAS Parallel Benchmarks*. page 79.
- [6] Bailey, D. H., Barszcz, E., Barton, J. T., Browning, D. S., Carter, R. L., Dagum, L., Fatoohi, R. A., Frederickson, P. O., Lasinski, T. A., Schreiber, R. S., Simon, H. D., Venkatakrishnan, V., and Weeratunga, S. K. (1991). *The nas parallel benchmarks—summary and preliminary results*.
- [7] Balasubramanian, N., Balasubramanian, A., and Venkataramani, A. (2009). Energy consumption in mobile phones: a measurement study and implications for network applications. In Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, pages 280–293.
- [8] Balouek, D., Carpen Amarie, A., Charrier, G., Desprez, F., Jeannot, E., Jeanvoine, E., Lèbre, A., Margery, D., Niclausse, N., Nussbaum, L., Richard, O., Pérez, C., Quesnel, F., Rohr, C., and Sarzyniec, L. (2013). Adding virtualization capabilities to the Grid'5000 testbed. In Cloud Computing and Services Science, volume 367 of Communications in Computer and Information Science. Springer.
- [9] Banerjee, A. and Roychoudhury, A. (2016). Automated re-factoring of android apps to enhance energy-efficiency. In Proceedings of the International Conference on Mobile Software Engineering and Systems, pages 139–150.
- [10] Baskar, P., Joseph, M. A., Narayanan, N., and Loya, R. B. (2013). Experimental investigation of oxygen enrichment on performance of twin cylinder diesel engine with variation of injection pressure. In 2013 International Conference on Energy Efficient Technologies for Sustainability, pages 682–687. IEEE.

- [11] Bedard, D., Lim, M. Y., Fowler, R., and Porterfield, A. (2010). *Powermon: Fine-grained and integrated power monitoring for commodity computer systems*. In Proceedings of the IEEE SoutheastCon 2010 (SoutheastCon), pages 479–484. IEEE.
- [12] Ben Asher, Y. and Rotem, N. (2009). *The effect of unrolling and inlining for Python bytecode optimizations*. In Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference on - SYSTOR '09, page 1, Haifa, Israel. ACM Press.
- [13] Benmoussa, K., Laaziri, M., Khoulji, S., Larbi, K. M., and Yamami, A. E. (2019). *A new model for the selection of web development frameworks: Application to PHP frameworks*. International Journal of Electrical and Computer Engineering (IJECE), 9(1):695–703.
- [14] Blackburn, S. M., Diwan, A., Hauswirth, M., Sweeney, P. F., Nelson Amaral, J., Tuma, P., Pankratius, V., Nystrom, N., Moret, P., Kalibera, T., and et al. (2012). *Evaluate collaboratory technical report: Can you trust your experimental results?*
- [15] Blackburn, S. M., Garner, R., Hoffman, C., Khan, A. M., McKinley, K. S., Bentzur, R., Diwan, A., Feinberg, D., Frampton, D., Guyer, S. Z., Hirzel, M., Hosking, A., Jump, M., Lee, H., Moss, J. E. B., Phansalkar, A., Stefanović, D., VanDrunen, T., von Dincklage, D., and Wiedermann, B. (2006). *The DaCapo benchmarks: Java benchmarking development and analysis*. In OOPSLA '06: Proceedings of the 21st annual ACM SIGPLAN conference on Object-Oriented Programming, Systems, Languages, and Applications, pages 169–190, New York, NY, USA. ACM Press.
- [16] Borkar, S. (2005). *Designing Reliable Systems from Unreliable Components: The Challenges of Transistor Variability and Degradation*. IEEE Micro, 25(6).
- [17] Breitbart, J., Weidendorfer, J., and Trinitis, C. (2015). *Case study on co-scheduling for hpc applications*. In 2015 44th International Conference on Parallel Processing Workshops, pages 277–285. IEEE.
- [18] Brown, A. W., Kaiser, K. A., and Allison, D. B. (2018). *Issues with data and analyses: Errors, underlying themes, and potential solutions*. Proceedings of the National Academy of Sciences, 115(11):2563–2570.
- [19] Bujnowski, G. and Smołka, J. (2020). *Java and kotlin code performance in selected web frameworks*. Journal of Computer Sciences Institute, 16:219–226.
- [20] Bukh, P. N. D. (1992). *The art of computer systems performance analysis, techniques for experimental design, measurement, simulation and modeling*.
- [21] Burtscher, M., Zecena, I., and Zong, Z. (2014). *Measuring gpu power with the k20 built-in sensor*. In Proceedings of Workshop on General Purpose Processing Using GPUs, pages 28–36.
- [Caldeira et al.] Caldeira, A. B., Grabowski, B., Haug, V., Kahle, M.-E., Laidlaw, A., Maciel, C. D., Sanchez, M., and Sung, S. Y. *Ibm power system s822*.
- [23] Caldeira, A. B., Grabowski, B., Haug, V., Kahle, M.-E., Maciel, C. D., and Sanchez, M. (2014). *Ibm power systems s814 and s824 technical overview and introduction*. IBM Redbook REDP-5097-00.

- [24] Calero, C., Mancebo Pavón, J., and García, F. (2021). Does maintainability relate to the energy consumption of software?
- [25] Cardenas, L. G., Gil, J. A., Domenech, J., Sahuquillo, J., and Pont, A. (2005). Performance comparison of a Web cache simulation framework. In 19th International Conference on Advanced Information Networking and Applications (AINA'05) Volume 1 (AINA Papers), volume 2, pages 281–284 vol.2.
- [26] Chamas, C. L., Cordeiro, D., and Eler, M. M. (2017). Comparing rest, soap, socket and grpc in computation offloading of mobile applications: An energy cost analysis. In 2017 IEEE 9th Latin-American Conference on Communications (LATINCOM), pages 1–6. IEEE.
- [27] Chasapis, D., Schulz, M., Casas, M., Ayguadé, E., Valero, M., Moretó, M., and Labarta, J. (2016). Runtime-Guided Mitigation of Manufacturing Variability in Power-Constrained Multi-Socket NUMA Nodes.
- [28] Chiba, T., Yoshimura, T., Horie, M., and Horii, H. (2018). Towards selecting best combination of sql-on-hadoop systems and jvms. In 2018 IEEE 11th International Conference on Cloud Computing (CLOUD), pages 245–252. IEEE.
- [29] Coles, H., Qin, Y., and Price, P. (2014). Comparing Server Energy Use and Efficiency Using Small Sample Sizes. Technical Report LBNL-6831E, 1163229.
- [30] Coley, G. (2012). Beaglebone rev a6 system reference manual. Obtenido, 4(23):2012.
- [31] Colmant, M., Rouvoy, R., Kurpicz, M., Sobe, A., Felber, P., and Seinturier, L. (2018a). The next 700 cpu power models. Journal of Systems and Software, 144:382–396.
- [32] Colmant, M., Rouvoy, R., Kurpicz, M., Sobe, A., Felber, P., and Seinturier, L. (2018b). The next 700 CPU power models. Journal of Systems and Software, 144.
- [33] Corral, L., Georgiev, A. B., Sillitti, A., and Succi, G. (2014). Method reallocation to reduce energy consumption: an implementation in android os. In Proceedings of the 29th Annual ACM Symposium on Applied Computing, pages 1213–1218.
- [34] Couto, M., Pereira, R., Ribeiro, F., Rua, R., and Saraiva, J. (2017). Towards a green ranking for programming languages. In Proceedings of the 21st Brazilian Symposium on Programming Languages, pages 1–8.
- [35] Crist, J. (2016). Dask Numba: Simple libraries for optimizing scientific python code. In 2016 IEEE International Conference on Big Data (Big Data), pages 2342–2343.
- [36] de Matos, F. F. S., Rego, P. A., and Trinta, F. A. M. (2021). An empirical study about the adoption of multi-language technique in computation offloading in a mobile cloud computing scenario. In CLOSER, pages 207–214.
- [37] Desrochers, S., Paradis, C., and Weaver, V. M. (2016). A validation of dram rapl power measurements. In Proceedings of the Second International Symposium on Memory Systems, MEMSYS '16, page 455–470, New York, NY, USA. Association for Computing Machinery.

- [38] Destefanis, G., Ortu, M., Porru, S., Swift, S., and Marchesi, M. (2016). *A Statistical Comparison of Java and Python Software Metric Properties*. In Proceedings of the 7th International Workshop on Emerging Trends in Software Metrics, WETSoM '16, pages 22–28, New York, NY, USA. ACM.
- [39] Echtler, F. and Häußler, M. (2018). *Open source, open science, and the replication crisis in hci. association for computing machinery, new york, ny, usa*, 1–8.
- [40] Eddie Antonio Santos, Carson McLean, Christophr Solinas, and Abram Hindle (2017). *How does docker affect energy consumption? Evaluating workloads in and out of Docker containers*. The journal of systems & Software.
- [41] Efron, B. (2000). *The bootstrap and modern statistics*. Journal of the American Statistical Association, 95(452).
- [42] El Mehdi Diouri, M., Gluck, O., Lefevre, L., and Mignot, J.-C. (2013). *Your cluster is not power homogeneous: Take care when designing green schedulers!*
- [43] Fahad, M., Shahid, A., Manumachu, R. R., and Lastovetsky, A. (2019). *A comparative study of methods for measurement of energy of computing*. Energies, 12(11):2204.
- [44] Fieni, G., Rouvoy, R., and Seinturier, L. (2020). *Smartwatts: Self-calibrating software-defined power meter for containers*. In 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID), pages 479–488. IEEE.
- [45] Fieni, G., Rouvoy, R., and Seiturier, L. (2021). *Selfwatts: On-the-fly selection of performance events to optimize software-defined power meters*. In 2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid), pages 324–333. IEEE.
- [46] Gajewski, M. and Zabierowski, W. (2019). *Analysis and Comparison of the Spring Framework and Play Framework Performance, Used to Create Web Applications in Java*. In 2019 IEEE XVth International Conference on the Perspective Technologies and Methods in MEMS Design (MEMSTECH), pages 170–173.
- [47] Ge, R., Feng, X., Song, S., Chang, H.-C., Li, D., and Cameron, K. W. (2009). *Powerpack: Energy profiling and analysis of high-performance systems and applications*. IEEE Transactions on Parallel and Distributed Systems, 21(5):658–671.
- [48] Goodman, S. N., Fanelli, D., and Ioannidis, J. P. (2016). *What does research reproducibility mean?* Science translational medicine, 8(341):341ps12–341ps12.
- [49] Grambow, M., Meusel, L., Wittern, E., and Bermbach, D. (2020). *Benchmarking microservice performance: A pattern-based approach*. In Proceedings of the 35th Annual ACM Symposium on Applied Computing, pages 232–241, Brno Czech Republic. ACM.
- [50] Guimarães, M., Saraiva, J., and Belo, O. (2016). *Some heuristic approaches for reducing energy consumption on database systems*. DBKDA 2016, page 59.
- [51] Hackenberg, D., Ilsche, T., Schöne, R., Molka, D., Schmidt, M., and Nagel, W. E. (2013). *Power measurement techniques on standard compute nodes: A quantitative comparison*. In 2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), pages 194–204. IEEE.

- [52] Hackenberg, D., Ilsche, T., Schuchart, J., Schöne, R., Nagel, W. E., Simon, M., and Georgiou, Y. (2014). *Hdeem: high definition energy efficiency monitoring*. In 2014 Energy Efficient Supercomputing Workshop, pages 1–10. IEEE.
- [53] Hackenberg, D., Schöne, R., Ilsche, T., Molka, D., Schuchart, J., and Geyer, R. (2015). *An energy efficiency feature survey of the intel haswell processor*. In 2015 IEEE international parallel and distributed processing symposium workshop, pages 896–904. IEEE.
- [54] Hammouda, A., Siegel, A. R., and Siegel, S. F. (2015). *Noise-Tolerant Explicit Stencil Computations for Nonuniform Process Execution Rates*. ACM Transactions on Parallel Computing, 2(1):1–33.
- [55] Hankins, T. L. (1986). *A debate over experiment: Leviathan and the air-pump. hobbes, boyle, and the experimental life. steven shapin and simon schaffer. including a translation of hobbes's dialogus physicus de natura aeris by simon schaffer*. princeton university press, princeton, nj, 1985. xiv, 442 pp., illus. 60. Science, 232(4753):1040–1042.
- [56] Hasan, S., King, Z., Hafiz, M., Sayagh, M., Adams, B., and Hindle, A. (2016). *Energy Profiles of Java Collections Classes*. In 2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE), pages 225–236.
- [57] He, S., Manns, G., Saunders, J., Wang, W., Pollock, L., and Soffa, M. L. (2019). *A statistics-based performance testing methodology for cloud applications*. In Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2019, pages 188–199, Tallinn, Estonia. ACM Press.
- [58] Heinrich, F., Carpen-Amarie, A., Degomme, A., Hunold, S., Legrand, A., Orgerie, A.-C., and Quinson, M. (2017). *Predicting the Performance and the Power Consumption of MPI Applications With SimGrid*.
- [59] Hirst, J. M., Miller, J. R., Kaplan, B. A., and Reed, D. D. (2013). *Watts up? pro ac power meter for automated energy recording*.
- [60] Ilsche, T., Hackenberg, D., Graul, S., Schone, R., and Schuchart, J. (2015). *Power measurements for compute nodes: Improving sampling rates, granularity and accuracy*. In 2015 Sixth International Green and Sustainable Computing Conference (IGSC), pages 1–8, Las Vegas, NV, USA. IEEE.
- [61] Inadomi, Y., Ueda, M., Kondo, M., Miyoshi, I., Patki, T., Inoue, K., Aoyagi, M., Rountree, B., Schulz, M., Lowenthal, D., Wada, Y., and Fukazawa, K. (2015). *Analyzing and mitigating the impact of manufacturing variability in power-constrained supercomputing*.
- [62] Islam, S., Noureddine, A., and Bashroush, R. (2016). *Measuring energy footprint of software features*. In 2016 IEEE 24th International Conference on Program Comprehension (ICPC), pages 1–4. IEEE.
- [63] Jagroep, E., Procaccianti, G., van der Werf, J. M., Brinkkemper, S., Blom, L., and van Vliet, R. (2017). *Energy efficiency on the product roadmap: an empirical study across releases of a software product*. Journal of Software: Evolution and process, 29(2):e1852.

- [64] Jagroep, E., van der Werf, J. M. E., Jansen, S., Ferreira, M., and Visser, J. (2015). Profiling energy profilers. In Proceedings of the 30th annual ACM symposium on applied computing, pages 2198–2203.
- [65] Joakim v Kisroski, Hansfreid Block, John Beckett, Cloyce Spradling, Klaus-Dieter Lange, and Samuel Kounev (2016). Variations in CPU Power Consumption.
- [66] Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.
- [67] Kalibera, T., Mole, M., Jones, R. E., and Vitek, J. (2012). A black-box approach to understanding concurrency in dacapo. In Leavens, G. T. and Dwyer, M. B., editors, Proceedings of the 27th Annual ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications, OOPSLA 2012, part of SPLASH 2012, Tucson, AZ, USA, October 21-25, 2012, pages 335–354. ACM.
- [68] Khan, K. N., Hirki, M., Niemi, T., Nurminen, J. K., and Ou, Z. (2018). Rapl in action: Experiences in using rapl for power measurements. *ACM Trans. Model. Perform. Eval. Comput. Syst.*, 3(2).
- [69] Kocher, P., Horn, J., Fogh, A., Genkin, D., Gruss, D., Haas, W., Hamburg, M., Lipp, M., Mangard, S., Prescher, T., Schwarz, M., and Yarom, Y. (2019). Spectre attacks: Exploiting speculative execution.
- [70] Koomey, J. et al. (2011). Growth in data center electricity use 2005 to 2010. A report by Analytical Press, completed at the request of The New York Times, 9(2011):161.
- [71] Kothari, N. and Bhattacharya, A. (2009). Joulemeter: Virtual machine power measurement and management. MSR Tech Report.
- [72] Kurpicz, M., Orgerie, A.-C., and Sobe, A. (2016). How much does a vm cost? energy-proportional accounting in vm-based environments. In 2016 24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP), pages 651–658. IEEE.
- [73] Laros, J. H., Pokorny, P., and DeBonis, D. (2013). Powerinsight-a commodity power measurement capability. In 2013 International Green Computing Conference Proceedings, pages 1–6. IEEE.
- [74] LeBeane, M., Ryoo, J. H., Panda, R., and John, L. K. (2015). Watt watcher: fine-grained power estimation for emerging workloads. In 2015 27th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD), pages 106–113. IEEE.
- [75] Lengauer, P., Bitto, V., Mössenböck, H., and Weninger, M. (2017). A comprehensive java benchmark study on memory and garbage collection behavior of dacapo, dacapo scala, and specjvm2008. In Binder, W., Cortellessa, V., Koziolek, A., Smirni, E., and Poess, M., editors, Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering, ICPE 2017, L’Aquila, Italy, April 22-26, 2017, pages 3–14. ACM.

- [76] Li, Y. and Schwiebert, L. (2016). *Boosting Python Performance on Intel Processors: A Case Study of Optimizing Music Recognition*. In 2016 6th Workshop on Python for High-Performance and Scientific Computing (PyHPC), pages 52–58.
- [77] Libič, P., Bulej, L., Horky, V., and Tůma, P. (2014). *On the limits of modeling generational garbage collector performance*. In Proceedings of the 5th ACM/SPEC International Conference on Performance Engineering, ICPE ’14, page 15–26, New York, NY, USA. Association for Computing Machinery.
- [78] Lilja, D. J. (2005). Measuring computer performance: a practitioner’s guide. *Cambridge university press*.
- [79] Lipp, M., Schwarz, M., Gruss, D., Prescher, T., Haas, W., Fogh, A., Horn, J., Mangard, S., Kocher, P., Genkin, D., Yarom, Y., and Hamburg, M. (2018). *Meltdown: Reading kernel memory from user space*.
- [80] Lovicott, D. (2009). *Thermal design of the dell™ poweredge™ t610™, r610™, and r710™ servers*. Round Rock, Texas.
- [81] Ma, H., Simon, D., Siarry, P., Yang, Z., and Fei, M. (2017). *Biogeography-based optimization: a 10-year review*. IEEE Transactions on Emerging Topics in Computational Intelligence, 1(5):391–407.
- [82] Manotas, I., Sahin, C., Clause, J., Pollock, L., and Winbladh, K. (2013). *Investigating the impacts of web servers on web application energy usage*. In 2013 2nd International Workshop on Green and Sustainable Software (GREENS), pages 16–23.
- [83] Marathe, A., Zhang, Y., Blanks, G., Kumbhare, N., Abdulla, G., and Rountree, B. (2017). *An empirical survey of performance and energy efficiency variation on Intel processors*.
- [84] Margery, D., Morel, E., Nussbaum, L., Richard, O., and Rohr, C. (2014). *Resources Description, Selection, Reservation and Verification on a Large-scale Testbed*.
- [85] McCreary, H.-Y., Broyles, M. A., Floyd, M. S., Geissler, A. J., Hartman, S. P., Rawson, F. L., Rosedahl, T. J., Rubio, J. C., and Ware, M. S. (2007). *Energyscale for ibm power6 microprocessor-based systems*. IBM Journal of Research and Development, 51(6):775–786.
- [86] Mirowski, P., Mathewson, K., Branch, B., Winters, T., Verhoeven, B., and Elfving, J. (2020). *Rosetta code: Improv in any language*. In Proceedings of the 11th International Conference on Computational Creativity, pages 115–122. Association for Computational Creativity.
- [87] Mishra, S. and Srivastava, S. (2021). *Web development frameworks and its performance analysis—a review*. Smart Computing, pages 337–343.
- [88] Mishra, S. K., Puthal, D., Sahoo, B., Jayaraman, P. P., Jun, S., Zomaya, A. Y., and Ranjan, R. (2018). *Energy-efficient vm-placement in cloud data center*. Sustainable computing: informatics and systems, 20:48–55.
- [89] Modzelewski, K. (2020). *Pyston v2: 20% faster Python*.

- [90] Mukherjee, T., Banerjee, A., Varsamopoulos, G., Gupta, S. K., and Rungta, S. (2009). Spatio-temporal thermal-aware job scheduling to minimize energy consumption in virtualized heterogeneous data centers. *Computer Networks*, 53(17):2888–2904.
- [91] Murri, R. (2013). Performance of Python runtimes on a non-numeric scientific code. page 6.
- [92] Mytkowicz, T., Diwan, A., Hauswirth, M., and Sweeney, P. F. (2009). Producing wrong data without doing anything obviously wrong! *ACM Sigplan Notices*, 44(3):265–276.
- [93] Nanz, S. and Furia, C. A. (2015). A comparative study of programming languages in rosetta code. In 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, volume 1, pages 778–788. IEEE.
- [94] Noureddine, A. (2022). Powerjoular and joularjx: Multi-platform software power monitoring tools. In 18th International Conference on Intelligent Environments.
- [95] Noureddine, A., Bourdon, A., Rouvoy, R., and Seinturier, L. (2012). A preliminary study of the impact of software engineering on GreenIT. In 2012 First International Workshop on Green and Sustainable Software (GREENS), pages 21–27.
- [96] Noureddine, A., Islam, S., and Bashroush, R. (2016). Jolinar: analysing the energy footprint of software applications. In Proceedings of the 25th International Symposium on Software Testing and Analysis, pages 445–448.
- [97] Noureddine, A., Rouvoy, R., and Seinturier, L. (2015). Monitoring energy hotspots in software. *Automated Software Engineering*, 22(3):291–332.
- [98] Oi, H. (2011). Power-performance analysis of jvm implementations. In ICIMU 2011: Proceedings of the 5th international Conference on Information Technology & Multimedia, pages 1–7. IEEE.
- [Oliveira et al.] Oliveira, W., Oliveira, R., Castor, F., Fernandes, B., and Pinto, G. Recommending Energy-Efficient Java Collections. page 11. Chapter 2: Collections- Introducing the main study field - categorizing them with the characteristics of each category .
- [100] O’Neil, E. J. (2008). Object/relational mapping 2008: hibernate and the entity data model (edm). In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 1351–1356.
- [101] Ourmani, Z., Belgaid, M. C., Rouvoy, R., Rust, P., Penhoat, J., and Seinturier, L. (2020). Taming energy consumption variations in systems benchmarking. In Proceedings of the ACM/SPEC International Conference on Performance Engineering, ICPE ’20, page 36–47, New York, NY, USA. Association for Computing Machinery.
- [Pang et al.] Pang, A., Anslow, C., and Noble, J. What Programming Languages Do Developers Use? a Theory of Static vs Dynamic Language Choice. page 9.
- [Pankiv] Pankiv, A. Concurrent benchmark system for web-frameworks on Python. page 26.

- [104] Patil, P. and Parmigiani, G. (2018). *Training replicable predictors in multiple studies*. Proceedings of the National Academy of Sciences, 115(11):2578–2583.
- [105] Peng, R. D. (2011). *Reproducible research in computational science*. Science, 334(6060):1226–1227.
- [106] Pereira, R., Carcao, T., Couto, M., Cunha, J., Fernandes, J. P., and Saraiva, J. (2017a). *Helping Programmers Improve the Energy Efficiency of Source Code*. In 2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C), pages 238–240, Buenos Aires, Argentina. IEEE. ## main idea
This paper presents a technique to spot the energy leaks inside a program using a spectrum technique called SPELL (Spectrum-based Energy Leak Localization)
How it works
We have a matrix ($n*m$) where n is the number of the tests and m is the number of components (like classes, methods, packages, etc.) after this we calculate the oracle (a vector that says which component is responsible for what)
Contributions
This paper helps developers to spot the red areas in their code and optimize the energy consumption. As an example it helped to reduce the energy consumption of a java application 50% faster and with 18% more efficiency.
- [107] Pereira, R., Couto, M., Ribeiro, F., Rua, R., Cunha, J., Fernandes, J. P., and Saraiva, J. (2017b). *Energy Efficiency Across Programming Languages: How Do Energy, Time, and Memory Relate?* In Proceedings of the 10th ACM SIGPLAN International Conference on Software Language Engineering, SLE 2017, pages 256–267, New York, NY, USA. ACM. this paper discuss the energy consumption through memory usage however we aim to see the relation between CPU usage and energy consumption rather than memory usage.
- [108] Philippot, O., Anglade, A., and Leboucq, T. (2014). *Characterization of the energy consumption of websites: Impact of website implementation on resource consumption*. In ICT for Sustainability 2014 (ICT4S-14), pages 171–178. Atlantis Press.
- [109] Pinto, G. and Castor, F. (2017). *Energy Efficiency: A New Concern for Application Software Developers*. Commun. ACM, 60(12):68–75. things to extend :
use micro benchmarks , and say that each one is used to measure a specific thing (memory intensive , cpu intensive , io intesive ..etc).
- [110] Prokopec, A., Rosà, A., Leopoldseder, D., Duboscq, G., Tuma, P., Studener, M., Bulej, L., Zheng, Y., Villazón, A., Simon, D., Würthinger, T., and Binder, W. (2019a). *Renaissance: Benchmarking suite for parallel applications on the jvm*. In Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, PLDI 2019, page 31–47, New York, NY, USA. Association for Computing Machinery.
- [111] Prokopec, A., Rosà, A., Leopoldseder, D., Duboscq, G., Tuma, P., Studener, M., Bulej, L., Zheng, Y., Villazón, A., Simon, D., Würthinger, T., and Binder, W. (2019b). *Renaissance: benchmarking suite for parallel applications on the JVM*. In PLDI, pages 31–47. ACM.
- [112] Redondo, J. M. and Ortín, F. (2015). *A Comprehensive Evaluation of Common Python Implementations*. IEEE Software, 32(4):76–84. benchmarks link
<http://www.reflection.uniovi.es/python>.

- [113] Ribic, H. and Liu, Y. D. (2016). *Aequitas: Coordinated energy management across parallel applications*. In Proceedings of the 2016 International Conference on Supercomputing, pages 1–12.
- [114] 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):1343–1372.
- [115] Simakov, N. A., Innus, M. D., Jones, M. D., White, J. P., Gallo, S. M., DeLeon, R. L., and FOPTurlani, T. R. (2018). *Effect of meltdown and spectre patches on the performance of HPC applications*. CoRR, abs/1801.04329.
- [116] Stodden, V., Seiler, J., and Ma, Z. (2018). *An empirical analysis of journal policy effectiveness for computational reproducibility*. Proceedings of the National Academy of Sciences, 115(11):2584–2589.
- [117] Tschanz, J., Kao, J., Narendra, S., Nair, R., Antoniadis, D., Chandrakasan, A., and De, V. (2002). *Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage*. IEEE Journal of Solid-State Circuits, 37(11).
- [118] van der Kouwe, E., Andriesse, D., Bos, H., Giuffrida, C., and Heiser, G. (2018). *Benchmarking Crimes: An Emerging Threat in Systems Security*. CoRR, abs/1801.02381.
- [119] van der Kouwe, E., Andriesse, D., Bos, H., Giuffrida, C., and Heiser, G. (2018). *Benchmarking Crimes: An Emerging Threat in Systems Security*. arXiv:1801.02381 [cs].
- [120] van Kessel, J., Taal, A., and Grossos, P. (2016). *Power efficiency of hypervisor-based virtualization versus container-based virtualization*. University of Amsterdam.
- [121] Vassamopoulos, G., Banerjee, A., and Gupta, S. K. S. (2009). *Energy Efficiency of Thermal-Aware Job Scheduling Algorithms under Various Cooling Models*. In Contemporary Computing, volume 40. Springer.
- [122] Vasan, A., Sivasubramaniam, A., Shimpi, V., Sivabalan, T., and Subbiah, R. (2010). *Worth their watts? - an empirical study of datacenter servers*.
- [123] von Kistowski, J., Block, H., Beckett, J., Spradling, C., Lange, K.-D., and Kounev, S. (2016). *Variations in cpu power consumption*. In Proceedings of the 7th ACM/SPEC on International Conference on Performance Engineering, pages 147–158.
- [124] Wang, X., Zhao, H., and Zhu, J. (1993). *Grpc: A communication cooperation mechanism in distributed systems*. ACM SIGOPS Operating Systems Review, 27(3):75–86.
- [125] Wang, Y., Nörterhäuser, D., Le Masson, S., and Menaud, J.-M. (2018). *Potential effects on server power metering and modeling*. Wireless Networks.
- [Wang et al.] Wang, Y., Nörterhäuser, D., Masson, S. L., and Menaud, J.-M. *Experimental Characterization of Variation in Power Consumption for Processors of Different generations*. page 10.
- [127] Yet, A. W. F. (2016). *Cross-language compiler benchmarking*.

- [128] Zar, J. H. (2005). *Spearman rank correlation*. Encyclopedia of biostatistics, 7.
- [129] Zimmerman, D. W. (1987). *Comparative power of student t test and mann-whitney u test for unequal sample sizes and variances*. The Journal of Experimental Education, 55(3):171–174.

