An Analysis of Energy Consumption of JavaScript Interpreters with Evolutionary Algorithm Workloads

Juan J. Merelo-Guervós\textsuperscript{a}\textsuperscript{1}, Mario García-Valdez\textsuperscript{a}\textsuperscript{2}\textsuperscript{b} and Pedro A. Castillo\textsuperscript{a}\textsuperscript{c}

\textsuperscript{1}Department of Computer Engineering, Automatics and Robotics, University of Granada, Granada, Spain
\textsuperscript{2}Department of Graduate Studies, National Technological Institute of Mexico, Tijuana, Mexico

Keywords: Green Computing, Metaheuristics, JavaScript, Energy-Aware Computing, Evolutionary Algorithms.

Abstract: What is known as energy-aware computing includes taking into account many different variables and parameters when designing an application, which makes it necessary to focus on a single one to obtain meaningful results. In this paper, we will look at the energy consumption of three different JavaScript interpreters: Bun, node and deno; given their different conceptual designs, we should expect different energy budgets for running the (roughly) same workload, operations related to evolutionary algorithms (EA), a population-based stochastic optimization algorithm. In this paper we will first test different tools to measure per-process energy consumption in a precise way, trying to find the one that gives the most accurate estimation; after choosing the tool by performing different experiments on a workload similar to the one carried out by EA, we will focus on EA-specific functions and operators and measure how much energy they consume for different problem sizes. From this, we will try to draw a conclusion on which JavaScript interpreter should be used in this kind of workloads if energy (or related expenses) has a limited budget.

1 INTRODUCTION

From a language designed in the nineties for simple browser widgets and client-side validations (Goodman et al., 2007; Flanagan, 1998), JavaScript is nowadays the language most widely used by developers in their GitHub repositories (O’Grady, 2022), occupying this position since 2014 (O’Grady, 2014), mainly because it is almost exclusively the language needed for front-end programming (competing only with app development languages, such as Swift or Kotlin, or languages transpiled to JavaScript, such as Dart), while at the same time being strong for full-stack development, with solid support for the back end, including application servers, middleware, and database programming. Other popularity indices, such as TIOBE\textsuperscript{a}, that take into account other factors besides lines of code production, currently (2023) rank it as the seventh, although it was also the most popular language in 2014. It can be claimed, then, that it is among the most popular, if not the most popular language in current development.

Due to its popularity and the fact that it has a continuously evolving standard (ECMA, 1999), traditionally, there have been different virtual machines (or interpreters) to run its programs. During the first years, browsers were the only running platform available; however, the introduction of Node.js running on the V8 JavaScript Engine (Tilkov and Vinoski, 2010) gave it the popularity it has today; this popularity, in turn, provoked new interpreters to spawn like deno (Doglio, ) (written in Rust) and bun (Tomar, 2022) programmed in the relatively unknown language Zig.

No wonder, then, that JavaScript is also a popular language for implementing metaheuristics, especially evolutionary algorithms (EA). EA (Eiben and Smith, 2015) are population-based stochastic optimization algorithms based on the representation of a problem as a (often binary) "chromosome", and evolution of population of these "chromosomes" by random change (via "genetic" operators, mutation and crossover) and survival of the fittest (evaluation of those chromosomes via a so-called "fitness" function, and selection and reproduction of those that achieve the best values). From the early implementations in the browser (Smith and Sugihara, 1996; González et al., 1999; Langdon, 2004), whole libraries (Re-
vas et al., 2014), through complete implementations geared towards volunteer computing (Merelo et al., 2016). However, one of the criticisms leveraged towards these implementations is the (possible) lack of speed when compared to other compiled languages (mainly Java, very popular in metaheuristics implementations, or C++).

This is why, since implementation matters (Merelo-Guervós et al., 2011), choosing the right interpreter is going to have a significant impact on the performance of any workload; if you decide to choose JavaScript for any reason (such as seamless client/server integration, or be able to run your algorithm either on the browser or from the command line if desired) knowing which VM delivers the best performance is essential, either from the scientific, or software engineering points of view.

At the same time, with the advent of the concept of green computing (Kurp, 2008), it becomes increasingly important to measure not only the raw wall clock performance (which was the focus of papers such as (Merelo-Guervós et al., 2017)), but also to achieve a certain level of performance with a certain amount of energy consumption, or else to minimize the consumption needed to run a certain workload. This will be the main focus of this paper; since the core of the different JS virtual machines is different, and are created with languages with different focus (Rust is focused on memory safety (Noseda et al., 2022), Zig based on simplicity and performance (Kelley, 2019)), different energy consumption should be expected. Since all three languages can (roughly) run the same, unmodified source code, what we intend with this paper is to advise on which JS interpreter might give the lowest power consumption, the maximum performance, or both, so that EA practitioners can target it for their development.

The rest of the paper follows this plan: next we will present the state of the art; then we will describe the experimental setup in Section 3; results will be presented next in Section 4, and we will end with a discussion of results, conclusions and future lines of work.

2 STATE OF THE ART

The power efficiency of CPUs (computations per kilowatt-hour) has doubled roughly every year and a half from 1946 to 2009 (Koomey et al., 2011), this improvement has been mainly a by-product of Moore’s law, the trend of chip manufacturers to decrease in half the size and distance between transistors every two years. Unfortunately, it is expected that physical limits of electronics will slow down this miniaturization in the near future. Nonetheless, energy efficiency is becoming the most important metric of performance and selling point in hardware development, and it is an important driver for current innovation. The challenge of building more power-efficient systems, can be addressed at the hardware and software levels. In the software level, developers focus their attention on the energy consumption of software, proposing optimizations for more energy-efficient algorithm implementations. Algorithm comparatives nowadays include power efficiency as a performance metric, these include encryption algorithms (Mota et al., 2017; Thakor et al., 2021), estimation models for machine learning applications (García-Martín et al., 2019) and genetic programming (GP) (Díaz-Alvarez et al., 2018), and code refactoring (Ournani et al., 2021). Since metaheuristics are so extensively used in machine learning applications, its interest in research has grown in parallel to its number of applications. Many papers focus on analyzing how certain metaheuristics parameters have an impact on energy consumption. Díaz-Alvarez et al. (Díaz-Alvarez et al., 2022) studies how the populations size of EAs influences power consumption. In an earlier work, centered on genetic algorithms (GAs) (Fernández de Vega et al., 2020), power-consumption of battery-powered devices was measured for various parameter configurations including chromosome and population sizes. The experiments used the OneMax and Trap function benchmark problems, and they concluded that execution time and energy consumption do not linearly correlate and there is a connection between the GA parameters and power consumption. In GAs, the mutation operator appears to be a power-hungry component according to Abdelhafez et al. (Abdelhafez et al., 2019), in their paper they also report that in a distributed evaluation setting, the communication scheme has a greater impact. Fernández de Vega et al. (de Vega et al., 2016) experimented with different parameters for a GP algorithm and concluded that handheld devices and single-board computers (SBCs) required an order of magnitude less energy to run the same algorithm.

3 METHODOLOGY AND EXPERIMENTAL SETUP

There are many ways to measure the consumption of applications running in a computer; besides measuring directly from the power intake, those running as applications and tapping the computer sensors fall roughly into two fields: power monitors and energy...
profilers (Cruz, 2021). Power monitors need additional hardware to measure the power drawn by the whole machine; besides being expensive, their setup is difficult, and it is complicated to measure precisely how much a specific process consumes.

On the other hand, energy profilers are programs that draw information from hardware sensors (Sinha and Chandrasakan, 2001), generally exposed through kernel calls or higher-level wrapper libraries, to pinpoint consumption by specific processes in a time period. Tools that give these measures, either with a graphic or a command line interface, have been available for some time already, and have become more popular lately. One of the mainstream processor architectures, Intel, includes an interface called RAPL, or Running Average Power Limit (Pandravada, 2014). Essentially, it consists of a series of machine-specific registers (MSRs) that contain information on the wattage drawn by different parts of the architecture; the content of these registers will be processed (through the corresponding library) and consumed by different command line utilities. We will use these command line utilities since they produce an output that can be automatically processed and evaluated, which is what we are looking for in this paper.

Energy profiling, as measured by RAPL or other APIs, includes different domains (Khan et al., 2015), essentially the computing devices or peripherals requiring the reported amount of energy. DRAM or dynamic RAM, CORE, or GPU will report what happens on those specific devices, with a core being every one of the computing units within the central processing unit; other domains, like PKG or package, will report what happens in the “package”, or CPU together with other devices in the chipset.

Considering that the available system has an AMD architecture, which is roughly compatible with the RAPL architecture, we will use two command line utilities, as they are the only ones available that can wrap the execution of a command and report on consumption for that specific command. These are open source tools that can be obtained directly from the corresponding repositories or by downloading and compiling their source code.

- pinpoint (Köhler et al., 2020) (available from https://github.com/osmhpi/pinpoint) is a tool that uses the RAPL interface, as well as the NVIDIA

- likwid-powermeter4 measures the CORE as well as the PKG domain, which includes the former together with the so-called “uncore” components.

On the JavaScript side, we used three different interpreters:

- bun version 0.5.8
- deno version 1.32.1, which includes the v8 library version 11.2.214.9 and typescript 5.0.2
- node.js version 18.5.0

bun and nodejs are fully compatible, so they run exactly the same code. The code for deno needed a small modification: the path to the library had to be changed (since it does not use the node_modules to host installed modules), and it uses a different library for processing the command line arguments. Other than that, the business logic was exactly the same.

These were running in an Ubuntu version 20.04.1 with kernel version 5.15.0-69. The processor is an AMD Ryzen 9 3950X 16-Core. Since we will not be testing in a pure Intel architecture, the complete RAPL API is not going to be available; that is also why we will be experimenting with different tools so that we can have an adequate coverage of energy consumption for the commands we will be measuring.

A Perl script was created to perform the experiments; it launched the scripts and collected results by analyzing the standard output and putting it in a CSV format that would allow examination of the experiments.

The initial experiment consisted in a script that used the saco-js library to perform the union of “bags”, sets that can hold several copies of the same item. 1024 sets were generated; these sets had 1024, 2048, 4096 elements. Then, a union of bags was performed on pairs of sets until there was only one left. This is similar to some operations performed by EAs, mainly related to merging populations. They do not involve floating point operations in any way.

---

2Systems based on the AMD architecture have a similar power profiling system called AMP with its corresponding command line tool. However, we found that it was not well documented, and excessively complicated for the purposes of this article. Although not as complete, AMD processors also include the aforementioned MSRs so RAPL-based utilities can run on them.

3https://perf.wiki.kernel.org/index.php/Main_Page

4https://github.com/RRZE-HPC/likwid
The scripts to launch every kind of tool were slightly different, mainly because the output needed to be processed in different ways (and different kinds of information extracted). Additionally, \textit{pinpoint}, which is the only tool that does not need superuser privileges, sometimes returned 0 in energy measures. This was an error, and those runs were discarded.

Finally, the scripts performed an additional task: since it is not possible to disaggregate the readings for our program from the energy consumed by other processes running at the same time, what we did was to run every program 15 times, compute the average time, and then use the same tool to measure the energy consumption for the \texttt{sleep} program during the average amount of time. The energy readings shown are the result of subtracting this measurement from every one of the 15 other measurements taken, so that we can analyze the differential of energy that has been consumed by our programs; the result is clipped at 0, since negative energy differentials would make no sense.

Experimenting with this simple program will mainly allow us to calibrate the different tools in order to pick only one, if possible, as well as validate the program and iron out all possible errors in the program itself or the processing scripts.

One of the first observations we can draw from these initial experiments is whether measure how much energy the GPU spends; it is interesting. Since \textit{pinpoint} is the only tool able to measure this, we will use it. We show the results in Figure \ref{fig:gpu}. We see here that it is mostly independent of the set size, but most importantly, it is 0 in most cases, more than 50%. The rest must be essentially noise, amounting to a few Joules anyway. In order to decide if we can use this tool alone or complement it with other measures, we need to validate or enhance its measurements with others; that is why next, we will make some test measurements with the next tool, \textit{likwid}.

Another tool that we have tested, \textit{likwid}, can also measure the \texttt{package} energy consumption. In this initial exploration, we will see whether this mea-

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1}
\caption{Boxplot of measurements of energy expenditures by the GPU in the sets problem. GPU energy consumption is measured in Joules.}
\label{fig:gpu}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2}
\caption{Boxplot of (differential) measurements of energy consumption as measured by the CORE and PKG (package) registers, both measured in Joules.}
\label{fig:core_pkg}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3}
\caption{Boxplot of (differential) measurements of energy consumption as measured by the CORE and PKG (package) registers, both measured in Joules.}
\label{fig:core_pkg}
\end{figure}
measurement is significant and to what extent it is related to the core measurements. Boxplots of energy consumption for different set size is shown in Figure 3 for the CORE (top) and PKG RAPL registers. It seems to show that differential energy consumption (once the baseline has been subtracted) is almost negligible in the CORE register unless the size is big enough (4096) in both cases, although the CORE register shows a certain amount of consumption for size = 2048. It is entirely unlikely that these kinds of measurements have a certain degree of uncertainty; however, apparently, pinpoint will directly estimate these runs as 0, and thus are skipped; however, this tool does measure a certain amount of consumption that cannot be so easily filtered.

Plotting the relationship between these two registers (see Figure 4), however, shows a linear relationship, so we do not really need to plot both. One of them will be enough, and since PKG seems to have the greatest variation, we will stick to that one.

Since several tools are available to measure PKG, we need to find out what kind of measurements they have, and if there is a correlation or even equality between them. We show how PKG energy consumption is registered by pinpoint and likwid in Figure 5. We have already seen in Figure 3 how the measurements registered by likwid have some strange behavior; this chart shows that while pinpoint shows a reasonable amount of consumption for all three sizes, likwid just registers zero, which is not reasonable in this case. This will lead us to discard the use of this tool in the upcoming experiments.

Unlike what happened with the previous tool, Figure 6 shows that there is a certain agreement between these two, except for the fact that perf seems to measure exceptionally low values in the lowest sizes. A Wilcox test shows significant differences for all three sizes; however, this might be due to a different amount of overhead or other unknown environmental factors.

Eventually, since pinpoint is able to measure all size ranges accurately, we will use it exclusively, as well as the quantity it measures, for comparison of different virtual machines. The initial results for this
Figure 7: PKG measurements for the set problem and the three different virtual machines.

Figure 8: Boxplot of energy consumption per second, in Joules/s.

Figure 9: Boxplot of energy consumption vs. time taken for all three sizes and VMs.

test drive are shown in Figure 7. This is due, in part, to the fact that they take a different amount of time, so it would be interesting to find out whether the energy density, or the consumption of energy per second, is similar. This is shown in Figure 8.

The interesting thing about this Figure is that the consumption per second varies with the VM used, as well as the size. Remarkably, deno keeps it approximately constant, while bun and node exhibit a variation with the size, and not in a systematic way. It is interesting also to note that while bun, in general, will spend less energy for each unit of work done than node, it will do so at the expense of exercising the CPU package more strenuously, spending much more energy per second than the other two VMs.

These, however, are preliminary findings with a test problem; we will have to design experiments for actual operations used in EAs, which we will do next.

4 EXPERIMENTAL RESULTS

As was done in (Merelo et al., 2016), which was focused on wallclock performance, the experiments will be focused on the key operations performed by an EA: evaluation of fitness and "genetic" operators like mutation and crossover. What we will do is, repeating the setup in the initial exploration, check the energy consumption for the processing of 40000 chromosomes, a number chosen to take a sizable amount of memory, but also on the ballpark of the usual number of operations in an EA benchmark, it is also small enough to not create garbage collection problems with the memory, something that was detected after the initial exploration. Experiments were repeated for the same chromosome size as before, 1024, 2048, and 4096, and for the three JS virtual machines used. Although the business logic is exactly the same for the experiments, the script run comes in two versions, one for deno and the other for bun/node, due to the different way they have of reading command-line arguments. This does not affect the overhead in any way. Code, as well as the data resulted from the experiments and analyzed in this paper, are released with a free license (along with this paper) from the reposi-
First, we will evaluate a typical fitness function, OneMax, which counts the number of ones in a binary (1s and 0s) string. This type of functions, which check the values of bits in a string and assign an integer value to it, are typical of many papers focused on evaluating EAs, including parallel versions (Merelo Guervós and Valdez, 2018).

The results are shown in Figure 9. This already shows that time, as well as energy consumption, for node is higher; just check the separation of the squares representing individual experiments in that interpreter to the rest of the values for the same color; this separation increases with chromosome size. But this paper focuses on energy consumption, which we summarize next in Figure 10.

The figure shows the almost-flat growth of energy consumption for bun. How consumption grows for deno is weird, since it takes less energy when the chromosome is bigger (4096). Once again, node is the bigger energy guzzler, consuming up to 3 times more than deno on average, and more than 6 times as much as bun. We will see how this is reflected in monetary terms, taking into account that the cost in Spain today is around 0.2€/kWh. This cost, shown in table 1, reaches almost one-hundredth of a euro for the most "expensive" VM, node; that gives you an idea of the kind of cost the algorithms have, and also how this cost decreases almost an order of magnitude if bun is used.

The crossover operation involves copy operations between strings, as well as creation of new strings. We will again generate 40K chromosomes and group them in pairs; these strings will be crossed by interchanging a random fragment from one to the other and back. The resulting pairs will be stored in an array, which is eventually printed. The result of every experiment is shown in an energy vs. wallclock time chart in Figure 11.

The scenario is remarkably similar to the one shown in Figure 9. In the two cases, bun achieves the top performance and lowest energy consumption, and node is the worst. Average energy consumption is shown as a boxplot in Figure 12.

Here we can see again the surprising fact that deno takes the same amount of energy, on average, as node for size 2048, in a similar case to what happened for OneMax (shown in Figure 10). The difference between the thriftiest, bun, and the heaviest consumer, node, is approximately three times, in this case, less than in the case of the OneMax fitness function.

Table 1: Estimated cost of the OneMax runs for every VM and size, in €-cents.

<table>
<thead>
<tr>
<th>size</th>
<th>VM</th>
<th>average</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td>bun</td>
<td>0.0010921</td>
<td>0.0000419</td>
</tr>
<tr>
<td>1024</td>
<td>deno</td>
<td>0.0015286</td>
<td>0.0000730</td>
</tr>
<tr>
<td>1024</td>
<td>node</td>
<td>0.0022507</td>
<td>0.0002090</td>
</tr>
<tr>
<td>2048</td>
<td>bun</td>
<td>0.0014913</td>
<td>0.0000353</td>
</tr>
<tr>
<td>2048</td>
<td>deno</td>
<td>0.0037011</td>
<td>0.0003269</td>
</tr>
<tr>
<td>2048</td>
<td>node</td>
<td>0.0036966</td>
<td>0.0003216</td>
</tr>
<tr>
<td>4096</td>
<td>bun</td>
<td>0.0016519</td>
<td>0.0001929</td>
</tr>
<tr>
<td>4096</td>
<td>deno</td>
<td>0.0031257</td>
<td>0.0000933</td>
</tr>
<tr>
<td>4096</td>
<td>node</td>
<td>0.0094820</td>
<td>0.0010450</td>
</tr>
</tbody>
</table>
functions, as well as the test function, are quite conclusive, we should not expect anything different from the mutation function, so we will leave the experiments at these two, and proceed to the conclusions.

5 CONCLUSIONS

In this paper, we set out to study the energy efficiency of different JavaScript interpreters in EA workloads, by studying how short scripts that are extensively used in implementations of these algorithms work and how these energy expenses scale with the chromosome size. We have developed a methodology to accurately measure per-process consumption; we have also adopted a tool, pinpoint, that is able to give good estimations of sensor readings, discarding experiments where that estimation was not adequate; this tool has been calibrated by comparing its readings with other tools, which were also evaluated for the same purpose and eventually discarded. We have also adopted a benchmark-based approach, similar to the one used to measure performance, so that consumption for specific operations can be pinpointed, discarding noise produced by an implementation of a complete algorithm; that is, a whole program includes different operations, applied in different proportion, that might cancel each other. Testing short code paths, as proposed by (Hähnel et al., 2012) makes easier to understand their individual contribution to the overall consumption of the algorithm, and eventually optimize their specific code, or the number of times they are applied in the algorithm.

During the exploratory data analysis, we have established that, in Linux machines, pinpoint can be profitably used to measure per-process energy consumption, as long as these measurements are repeated and the process themselves include short snippets of business logic; this tool should be preferred over others that are either less accurate or simply take into account different aspects of energy consumption.

The main point of this paper, however, was to check which JavaScript interpreter should be used if our objective is to consume the minimum amount of energy; the experiments have reliably confirmed bun to be that tool. Not only it consumes less for all the range of chromosome sizes used; it also takes less time and can run applications written for Node (mostly) unmodified; its consumption also scales better with problem size. This might be due to design considerations, but also to the fact that the language used to write it, Zig, emphasizes compile-time safety and manual memory allocation by default, and avoids hidden control flow. The only inconvenience of this interpreter is that it has not reached version 1.0 yet, being currently in version 0.5.9; this might prevent any company or organization from using it in production environments.

If that is an issue, deno might be a good alternative. Except in a specific case, it is going to be faster and consume less energy than node.js, even more so when memory requirements are high. According to our initial exploration, it will also consume less energy per second, thus for workloads that take roughly the same time, it will be a better candidate than node.js. As an inconvenience, it needs minor modifications to run, at least if you need core or other kind of external libraries; its core library modules are different to those used in node/bun, although that need not be a disadvantage per se.

The previous two points imply that, energy-wise, there are no good reasons to use node.js for running EAs. Except if the business logic uses specific, early-adoption, or some features that, for some reason, does not work with bun yet, we would advise anyone to keep using bun for this kind of workloads.

As we have indicated in the experimental session, the wide advantage that bun has over the other interpreters does not leave much room for adopting different benchmarks that could make that ranking vary; at any rate, these experiments have shown how much faster and energy-saving bun is (from 1/3 to 1/6 the energy consumed by node.js), but it would be interesting to know what happens to this gap under different operations like selection or different kind of mutation. At the same time, even if EAs are mostly GPU-free, there are some fitness functions that operate on float-
ing point numbers and thus would need to use the GPU; how interpreters work in this area could be an interesting future line of work. This, along with testing different versions of the interpreters as they are published, will be the subject of future research.

ACKNOWLEDGEMENTS

This work is supported by the Ministerio español de Economía y Competitividad (Spanish Ministry of Competitiveness and Economy) under project PID2020-115570GB-C22 (DemocratAI:UGR).

REFERENCES


Doglio, F. Introducing Deno.


