Performance Implications of Programming Languages

Jason H. Sauntry Northeastern University
April 27, 2022 *Final Report* EECE 5640

In 2020, I did a co-op at Purefacts Financial Solutions in our machine learning R&D team. One problem we spent a lot of time talking about was the memory footprint of our algorithm, which was so big that in it's first test with a real client's data, 64 G of ram hadn't been enoguh, and the program didn't work. Because of this we spent a lot of time optimizing our code. And yet, the code was in Python, a language well-known for being rather inefficient. We made our program work, but I always wondered how much performance we lost by our choice of programming language.

The question I wanted to answer in my project is a generalization of the above: how much does one's choice of programming language affect performance? Everyone knows that C is faster than Python, but how much faster? And where do other languages fit in?

Specifically, I measured the following langauges:

- Python
- C
- Java

Using the following benchmarks:

- Cannonical Matrix Multiplication
- Optimized Matrix Multiplication
- Sorting (via standard library)
- Heap (measured by sorting a list)

Performance was measured by the following metrics:

- Runtime (wall-clock)
- Peak memory usage

1 Testing Methodology

I evaluated each language over a series of benchmarks, intended to represent a variety of general-purpose yet performance-intensive computing workloads.

The source code of all benchmarks are availaible on GitHub, along with instructions describing how to run them.

1.1 Choice of Benchmarks

1.1.1 Cannonical Matrix Multiplication

Matrix multiplication is a common benchmark in high-performance computing and is a useful task in many applications. It's memory-intensive but has somewhat good caching perfermance due to the storage of data in 2d arrays.

I implemented this multiplication myself, based on the following psuedocode:

```
1     matmul(matrix a, matrix b, matrix c) {
2         for (i = 0; i < N; i++) {
3             for (j = 0; j < N; j++) {
4                 c[i][j] = 0;
5                 for (k = 0; j < N; k++) {
6                     c[i][j] += a[i][k] * b[k][j];
7                 }
8             }
9         }
10    }
```

No manual optimizations were applied, but C code was be compiled with `-O3`.

1.1.2 Optimized Matrix Multiplication

In addition to seeing what a language can do in typical use, I was also interested in what it can do when it has been optimized to the highest extent possible. I again used matrix multiplication for this benchmark, since it's common enough to have highly-optimized libraries in each language I used. Specifically:

Python Numpy

C OpenBLAS

Java EJML

1.1.3 Sorting

Another common task in software is sorting a list. This is a performance-intensive task, and is included in the standard libraries of each language I am evaluating. This allowed me to evaluate the efficiency of standard library functions (as opposed to 3rd-party libraries evaluated in matrix multiplication).

1.1.4 Heap

The above benchmarks all use data organized in arrays. Arrays are good for caching since they have high spatial locality. But I was also interested in what happens when operating

on a pointer-heavy data structure, where closely related blocks of data may be far apart in physical memory address. I did this by implementing a heap as a binary tree.

To exercise this data structure, I sorted a list. I generated a list of floats, inserted each into the heap, then popped them all into a new list.

1.2 Data Extraction

There were two data points collected for exery benchmark: runtime and peak memory usage. Runtime was memasured directly by the language under test. Memory usage, on the other hand, was measured using Valgrind's Massif tool.

In both cases benchmarks were managed—and data collected—using a Python script. Another Python script generated graphs for analysis.

2 Results

2.1 Cannonical Matrix Multiplication

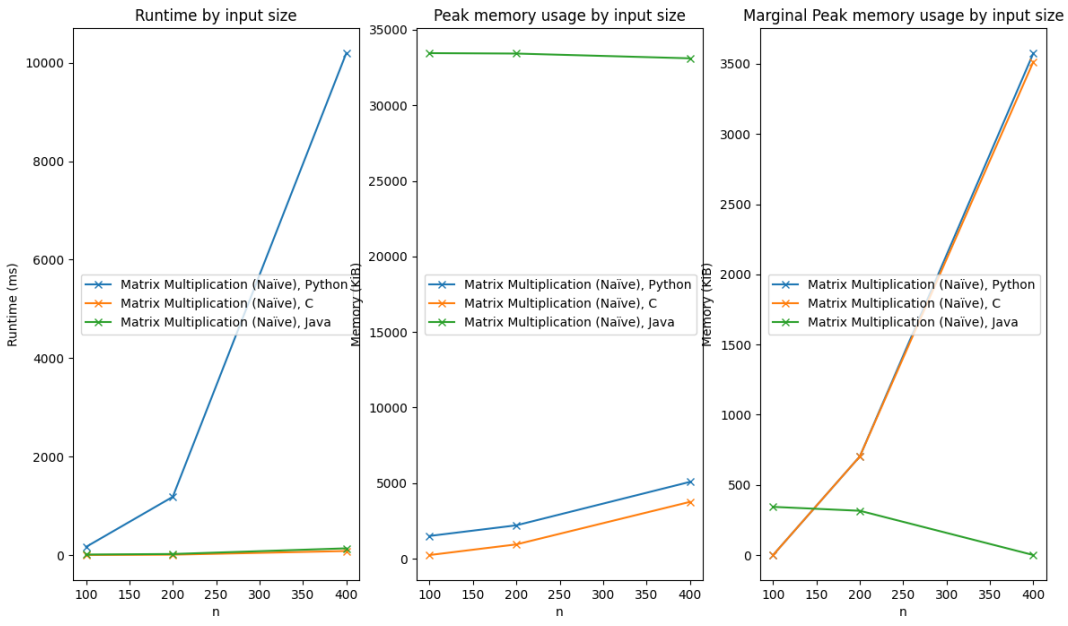

Figure 1: Performance of cannonical matrix multiplication.

Figure 1 shows the performance of cannoncical matrix multiplication. The leftmost graph

shows the runtime of the benchmark. This *only* measures the benchmark itself, not any setup or tear-down. In this case, runtime includes multiplying two matrices, but not allocating the input matrices and filling them with random data.

The center graph shows peak memory usage, measured using Valgrind. This measurement includes the entire benchmark program, including both setup and tear-down. This is because there is no good way in Valgrind to determine what part of the program is running at any given time. The rightmost graph shows *marginal* memory usage, which is the peak memory usage minus the lowest peak memory usage of that program and benchmark, for any input. It is intented to show the memory associated with increased input size, without any constant overhead.

C did very well in this benchmark. At all input sizes, it was both the fastest and used the least memory.

Python did very poorly on speed, taking 10 times the runtime as C on larger input sizes. It performed much better on memory. Although Python carried a 1 MB overhead, this overhead was constant. In marginal memory usage Python was very slightly worse than C.

Java on the other hand performed very well in runtime, with runtimes only doubling that of C on larger inputs. Java's limitation was in terms of memory, with performance *considerable* worth than Python or C. This memory usage was an overhead. Java's marginal memory usage appears very good, but that is probably due to random variance in the memory overhead making the marginal memory usage an invalid measurement.

2.2 Heap

Figure 2 shows the performance of my heap implementation. As with 2.1 Cannonical Matrix Multiplication, C's performance was boring and efficient.

Python's performance was similar to 2.1 Cannonical Matrix Multiplication: memory was efficiency, constant overhead notwithstanding, but runtime was very poor.

Java's performance[1] was interesting: the runtime of the sort was so low that it rounded to zero. I am dubious of the validity of this result, but as far as I can tell it's legitimate. The runtime is measured using the same function as

2.3 Standard Library Sorting

Figure 3 shows the performance of sorting via the various language's standard library. As usual, C was fast and efficient.

Python's performance on this benchmark was interesting. The runtime performance was about double that of C. That doesn't seem impressive—it's still slower—but in previous

[1] Java's results for this benchmark are different that what I presented in class on April 22, 2022. I noticed a bug after the presentation where the Java implementation skipped all the computation.

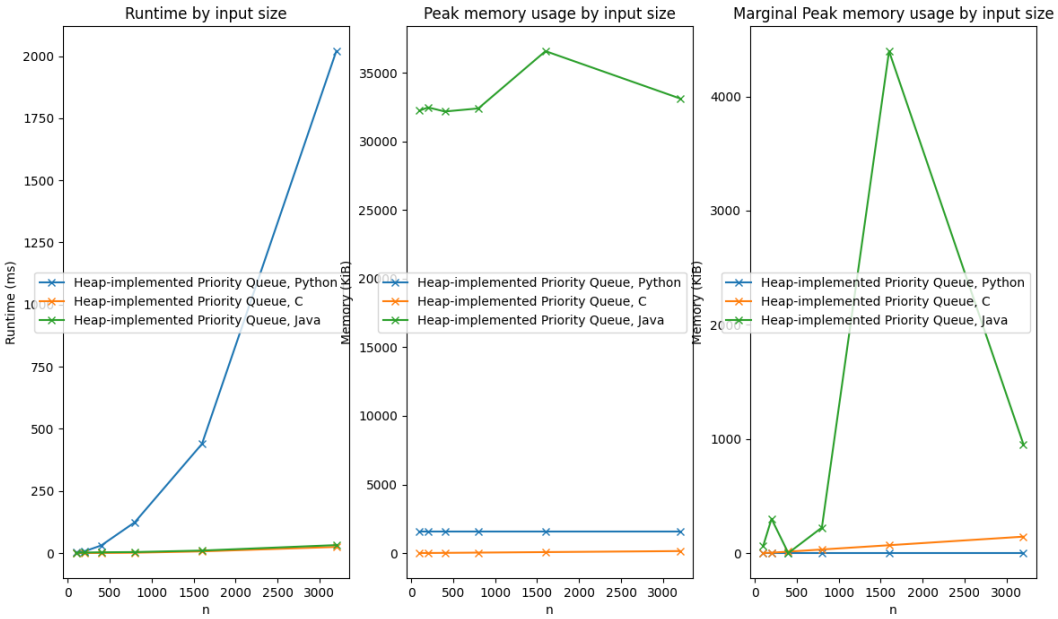

Figure 2: Performance of cannonical matrix multiplication.

benchmarks it was orders of magnitude slower. As usual, Python's memory performance was quite good. In fact, it was slightly better than C. There are two main reasons for this: firstly, the memory requirements of storing the input are high enough that the constant overhead is not noticable. This is because Python's sort implementation is in-place[2], whereas C's implementation may not be in-place[3] and thus may require additional heap space.

Java's runtime performance was similar to how the language behaved previously: worse than C, but not drastically so. It's memory performance on the other hand was impossibly good. This is discussed in **??**.

2.4 Optimized Matrix Multiplication

Figure 4 shows the performance of sorting via heavily-optimized third-party libries. This benchmark is intended to show what the language can do when it is optimized to the greatest extend possible.

C's matrix multiplicaiton was done using the OpenBLAS[4] library. As usual, C did reasonably well, but it was *not* the fastest in all cases.

[2]Per Python's documentation.
[3]In the GNU specification.
[4]Basic Linear Algebra System

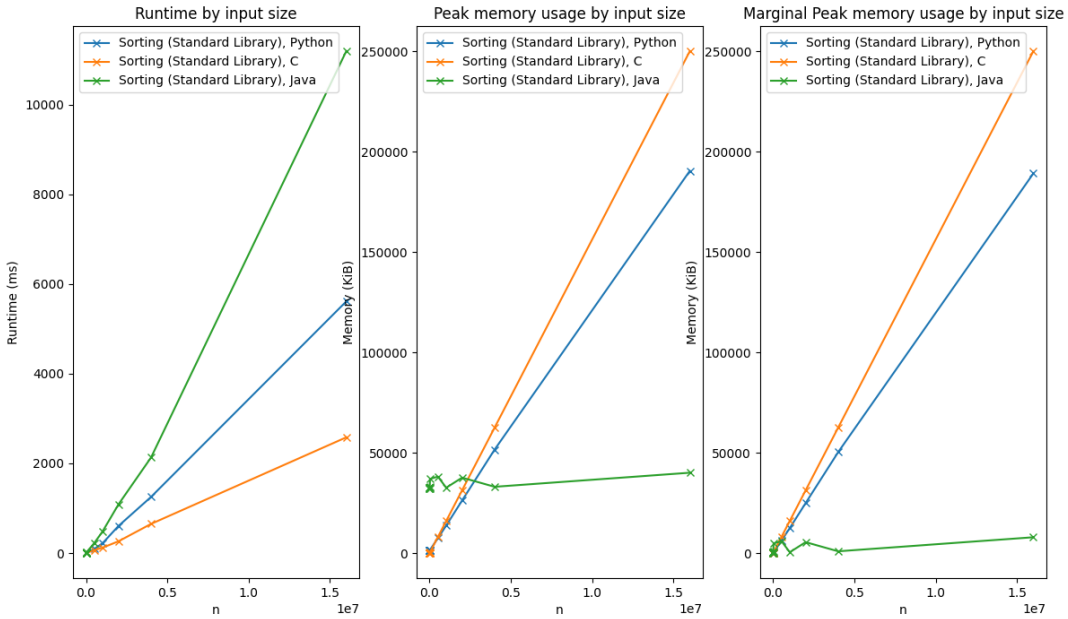

Figure 3: Performance of sorting via standard library functions.

Python used the popular Numpy library, although as discussed in **??**, this isn't *quite* correct. Impressively, Python's performance at the largest input size was slightly that C. It's marginal memory performance was also very good, almost exactly equal to that of C[5].

Java used the EJML linear algebra library. It's performance was unimpressive, with runtimes larger than C, but within the same order of magnitude.

3 Analysis

3.1 Why Python is so Slow

Something that's very clear from 2.2 Heap and 2.1 Cannonical Matrix Multiplication is that Python code is really, *really* slow.

On a high level, the reasons for this is rather simple. C and Java are both ahead-of-time compiled languages: transformantion of textual code to computer-friently bytecode is done before runtime, by a compiler (either GCC or the Java Compiler). Python, on the other hand, is a JIT langauge, or just-in-time compiled. Compilation is done *at* runtime. This

[5]This is why the Python line isn't visible on the marginal memory usage graph: The Python line is exactly underneath the C line.

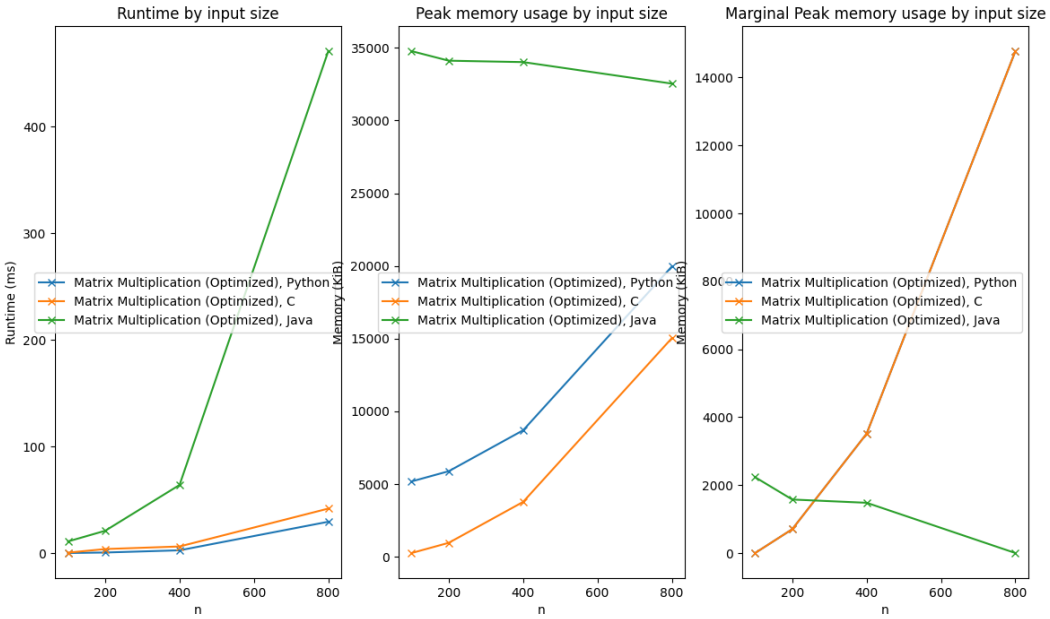

Figure 4: Performance of sorting via standard library functions.

adds a time overhead.

Python is also an interpreted language. C's bytecode is run directly on the CPU's hardware, with no layer of indirection.[6] But Python runs in the Python Interpreter, a program that interprets and runs your code for you. Need to reference a variable? Your code has to ask the interpreter to do it for you. Want to call a function? Again, the interpret adds a (slow) layer in between your code and a function call on the CPU. All this adds up to a significant slowdown.

Python isn't the only interpreted language in my sample: Java is also interpreted: it runs in the Java Virtual Machine (JVM). But the JVM is more efficient than the Python Interpreter, probably because it's a thinner abstraction.

3.2 Java Memory Overhead

It's very apparent from all benchmarks that before allocating *any* significant data, Java has a large memory overhead, consistently about 30 MB. This is because of the JVM. As it turns out, the JVM has to store a lot of data about the code itself: compiled classes, field names and function signatures, as well as the garbage collector. In particular, compressed class

[6] Except for the indirection built into CPU itself, but that is common to all programs.

space grows proportionally to the number of classes loaded[7], and can easily dwarf the heap.

For more information, see Memory footprint of the JVM, an article by Spring Blog that delves into this topic in some detail. For my purposes, the main takeaway is that every Java process that gets launched will immediatly take up 30 MB of RAM.

The question I'm more interested in is this: is this a significant loss of available memory? Well, it depends. Certainly it's problematic in embedded applications where available memory is sometimes measured in bytes. But Java is not commonly used for embedded tasks. Java is normally run on modern machines like telephones and computers, where gigabytes of memory are available. Suppose an application is being developed that will run on a server with 128 GB of ram available. Is this 30 MB overhead significant? Again, it depends. If the deployment plan is to run a handful of processes that each do a lot of work, the memory overhead probably isn't significant. If the plan is to run microservices, with thousands of process each doing a tiny piece of work, then this tiny overhead is now taking up gigabytes of memory, and that probably *is* a problem.

3.3 Why Sorting in Python *isn't* Terrible

As noted in 3.1, Python is rather painfully slow. Except, in 2.3 Sorting[8], it isn't. Why? Because the sorting isn't done in Python code. It's done directly by the Python Interpreter. Or, to be more precise, it's done by the CPython interpreter, since CPython is the Python implementation under test. CPython is implemented in C. So Python's sorting is implemented in C[9][10] which, as this project has made abaundantly clear, is quite fast.

So speaking generally, what are the implications of this? It means that although Python code runs slowly, standard library functions run quite quickly. So will a performance-intensive application run in a reasonable time in Python? It depends. If most of the computation happens within a standard library function, it might run quite quickly.

3.4 Why Python's Numpy is Fast

So far, we've established that Python code is very slow, but Python standard library functions can be fast, though not *as* fast as C. So one would expect that Python's optimized matrix multiplication runs very slowly. It's certainly not a standard library function. And yet, Numpy's performance was approximately the same as C, both in terms of runtime and in terms of marginal memory usage.

How is this possible? Well, Python (or, at least, the popular CPython implementation) has a rarely-used but enormously powerful feature: it can interface nicely with C and C++

[7] Which in turn is roughly proportional to the complexity of your code.
[8] It also wasn't slow in 2.4. I'll address this in **??**.
[9] For reference, see the header sorting where sorting is declared and the implementation.
[10] Presumably, if I tested Jython, it would be implemented in Java.

code[11]. This means that a Python program with a performance-intensive component can implement that component in C, taking advantage of C's efficiency. This is exactly how Numpy implements matrix multiplication: in C. In fact, it uses OpenBLAS. Most of the work in the Python matrix multiplication benchmark happens in a call to `cblas_dgemm`, the exact same function used in my C benchmark. So of *course* Python's performance was similar to C's: on some level, it was running exactly the same code.

4 Conclusion

Before starting this project, I had a vague idea that C was very efficient and Python was very slow, with most languages in the middle. In other words, to program in a easy-to-use language like Python, you have to pay the price of worse performance. My goal in this project was to quantify this cost, with the ultimate goal of helping answer the qustion of whether that cost is worth it.

To absolutely no one's surprise, the answer is that it depends.

The performance penalty will depend not only on what language is used, but also on *how* that language is used. Python is the most extreme example of this. A pure-Python implementation might be very slow. But perhaps an application is 90 % Python, 10 % C? If the performance-intensive part is in C, you can sort of get the best of both worlds: most of your code is built in the easy-to-use Python, but code that has to run fast is built in the more efficient C.

Ultimately, every language will have advantages and disadvantages. To intelligently choose the best language for a particular task, one needs to know what these advantages and disadvantages are. To that end, here are listed the specific, objective results of this project:

- The runtime of Python code can be over 100 times the runtime of equivalent C code.
- The runtime of Java code is generally about double the runtime of equivalent C code.
- Each Python process introduces a memory overhead of approximately 100 kB.
- Each Java process introduces a memory overhead of approximately 30 MB.
- The marginal cost of allocating more data in Python is not materially different than allocating equivalent data in C.
- The marginal cost of allocating more data in Java is difficult to measure.
- Python standard library functions can be nearly as fast or as fast as their C equivalents. The same is true of third-party libraries that have implemented some of their functionality in C.

[11] Extending Python with C or C++

- CPython allows easy integration with C code, meaning that a Python codebase can, in some cases, be easily accelerated by implementing a small portion of the codebase in C.