**21/12/2020**

**Substring matching with application to genomics/proteomics**

**Report by:**

**Mladen Jurkic – Chiara Balduzzi**

**Introduction:**

The aim of the developed algorithm is to find a group of characters (called ‘*pattern’*) among a larger set (indicated as *‘sequence’*). Both sets are only composed out of 4 elementary characters which refer to a nitrogenous base that could be found inside a nucleic acid chain (i.e.: DNA): adenine (A), thymine (T), guanosine (G), cytosine (C).

In our program, the comparison between *pattern* and *sequence* can be made in both forward and backward directions. Assuming the length of the sequence and of the pattern to be, respectively, *‘slen’* and *‘plen’*, we can have a total or partial overlapping.

In this latter case, a match is found when the sub-sequence and pattern are equal up to a limited number of contiguous errors (denoted as *‘holes’* – at most 10% of *plen*) having a maximum allowed width, which in our implementations is set to 0.1% of the length *plen*. These two percentages come from a trial-and-error kind of approach and seemed reasonable to us.  
In either case it is possible to get the positions of the sequence at which the correspondence was found. Obviously, by adding the possibility to take holes into account we increase complexity and overall computational time of the program.

A first idea to solve the given problem, could be to use the built-in C function strcmp: int  
This function can compare the *pattern* with a sub-string of the *sequence,* both having the same length *‘plen’*. By selecting, with proper indices, the sub-string of the *sequence* and ending the loop when we reach ‘*slen-plen’*, it’s possible to obtain the number of perfect correspondences. However, we abandon this solution favouring a “manual” approach for two main reasons.

First, by using the built-in function we cannot appreciate the effect of the parallelization as the function is already optimized and very fast. Furthermore, the need to include partial overlaps makes this solution not useful for us.

To implement the ‘perfect-match’ solution we had to compare single characters of the sequence, one at a time: in case of a match of a character, we proceed in the comparison of the successive ones and obtain a correspondence if all of them match. Otherwise, we simply increment a counter variable, shifting the beginning of the sequence and de facto comparing the pattern with a new sub-sequence.  
The same reasoning lies behind the partial-match solution, where in addition we have an error counter to keep track of the holes. The comparison is abandoned if the number of holes or their width is greater than a given threshold.

**Preliminary analysis of the algorithm:**

In the serial code development, the first idea was that better performances could be obtained by writing a single function that is able to scan the sequence in a forward direction (called *‘search\_fwd’*).

The backward scanning can simply be obtained reversing the pattern and inspecting the sequence forward once more. This solution is presented in the file *‘serial\_v1.c’* and has the advantage of re-utilizing the same code. What is more, the pattern reversal time is practically irrelevant, even for large patterns. However, having in mind that our goal is to implement a parallel version of the code, we provided another version (*‘serial\_v2.c’*) in which a single function sweeps the sequence both in forward and backward direction.

Our hope was to be able to better parallelize this version by having some threads working to scan in one direction and some others in the opposite direction.

The common features of the two versions are the main function where we ask the user to insert the desired length of both sequence and pattern and the *‘generate’* function that randomly generates the two sequences. We also provided a version of the codes where we do not generate the sequences but read them from file instead, if the user wishes to do so (same names as before with *‘\_file’* suffix).

Running the codes, we can analyse the difference in timings of the two versions taken using *omp\_get\_wtime()*: results can be seen in Table 1. For simplicity, the string generation time was omitted as it was quite constant throughout the different combinations of sequence and pattern lengths and can be even smaller and more negligible if we read from file. The analysis was performed without reading from file.

What we can notice is that having a fixed pattern length, the increase of *‘slen’* provides a slightly better time and fraction of parallelizable code. Instead, keeping a constant sequence length and enlarging the pattern, the improvement obtained is much greater.

Below, we can also appreciate in Figure 1 the difference between match and total time, that is to say the behaviour of the percentage of parallelizable code, of version V1 in the serial implementation if we increment the two strings’ size. The serial V2 plot is similar to the first version.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SLEN | PLEN | MATCH TIME | TOTAL TIME | Parallelizable % | MATCH TIME | TOTAL TIME | Parallelizable % |
|  | | **Version 1** | | | **Version 2** | | |
|  | **10** | 0.00003 | 0.0001 | 29.62188 | 0.00003 | 0.00012 | 25.05513 |
|  | **10** | 0.0002 | 0.00046 | 43.478 | 0.00021 | 0.00043 | 48.96214 |
|  |  | 0.00019 | 0.00038 | 51.31718 | 0.00021 | 0.00040 | 51.26549 |
|  | **10** | 0.00194 | 0.00305 | 63.73054 | 0.00211 | 0.00351 | 60.12327 |
|  |  | 0.00177 | 0.0029 | 60.88906 | 0.00192 | 0.00315 | 61.15258 |
|  |  | 0.00448 | 0.00581 | 77.00919 | 0.00310 | 0.00443 | 69.90329 |
|  |  | 0.01949 | 0.02967 | 65.68999 | 0.01969 | 0.03111 | 63.26750 |
|  |  | 0.04122 | 0.05085 | 81.06793 | 0.03432 | 0.04578 | 74.95414 |
|  |  | 0.19065 | 0.28348 | 67.25468 | 0.19736 | 0.30575 | 64.54814 |
|  |  | 0.44433 | 0.53645 | 82.82867 | 0.34026 | 0.44850 | 75.86705 |
|  |  | 12.81745 | 12.90995 | 99.2835 | 13.10603 | 13.19893 | 99.29617 |
|  |  | 4.33195 | 5.23731 | 82.71319 | 3.26960 | 4.34708 | 75.21382 |
|  |  | 137.68859 | 138.59632 | 99.34505 | 144.16187 | 145.08619 | 99.36292 |

Table 1: comparison between serial V1 and serial V2.

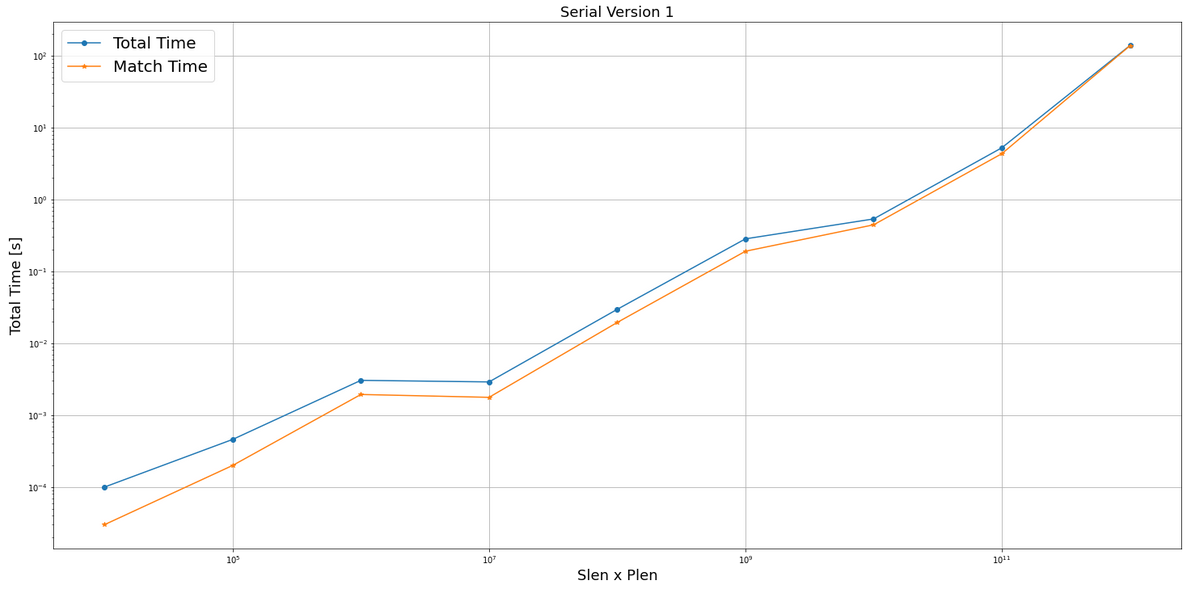


Figure 1: serial V1. Increment of percentage for greater sequence and patterns.

As a result of this analysis, we can apply Amdahl’s law to predict what is the expected speedup in the parallel code.  
Chosen a number of cores *n* and using the percentage of parallelizable code *f* obtained in the previous table we expect a speedup equal to: .

In a practical case, having 16 cores and the combination *slen-plen* of we expect a speedup of 14.45 in the version 1, which is rather promising.

**Parallel OpenMP implementation:**

To get to a parallel version, first of all we need to choose the number of threads that will be working synchronously. Thus, it was made possible for the user to specify the number of threads (accordingly to the type of architecture the code is run onto). This is then set using the OpenMP library function *omp\_set\_num\_threads(n\_thr)*.  
As a choice, we did not parallelize the *‘generate’* function as it is a random process of string generation. Instead, we put our efforts on the main core of the whole program, that is the *‘search’* function. Being it a set of two *for* loops, a first and very simple approach could be to use the following OpenMP construct: *#pragma omp parallel for.*

in which we obtained the best results selecting a dynamic type of schedule, where each idle thread is given the first available chunk. In order to manage load balancing and to reduce overhead, we decided to increase the chunk size to the ratio between *‘plen’* and the number of threads. Furthermore, we had to insert two critical sections: one is needed to synchronize the update of the counter of correspondences and of their positions. The other one is used to reallocate memory, using the *realloc* C function, in case the correspondence vector gets bigger than what we expected. Clearly, these two critical sections have an impact on execution times as they are synchronization points where threads might need to wait before modifying shared variables. This reasoning is reflected into our *‘parallel\_v1.c’* code.

Another feasible solution could be to implement tasks to parallelize. To do so, we exploited the fact that our version 2 of the serial code had a *search* function that performed both forward and backward in the same function. This new parallel version, indicated as *‘parallel\_v2.c’*, does not need the setting of the number of threads as now we are using tasks. The code contains the ‘*sequencing’* function in which we put a parallel section obtained via an OpenMP pragma.

Inside this parallel region, we declare, again using a pragma, two different tasks having *first-private* error counters (initialized to zero): one for the forward search and the other one for the backward scanning. In order to avoid the tasks to be spawned more than once inside the parallel region, we needed to introduce a *#pragma omp single.* Finally, inside the task region we could write our ‘*parallel for’* region that sweeps the two different loops for the two different scanning directions while maintaining the critical sections that we introduced in the first version, too. It is rather obvious that this new version has a much greater complexity and a deeper level of parallelization. It is then possible that we will have to keep in mind a trade-off among complexity and performance.

**Testing:**

After the development of the software and some basic testing to check the correspondences matched, we switched to the “real” testing phase. We greatly enlarged the sizes of both sequence and pattern and started analysing the serial code, first, and then the parallel versions where we could try to improve the results using more cores. To get the needed computational power, we could not use our personal computers and, thus, had to access the VMs provided by Google Cloud Platform.  
Starting from the parallel version ‘V1’ we obtained quite varying timings for smaller sequence/pattern combinations, while for larger numbers the outcomes were quite similar. In both cases the provided times are an average of more executions. The results are shown in Table 2.

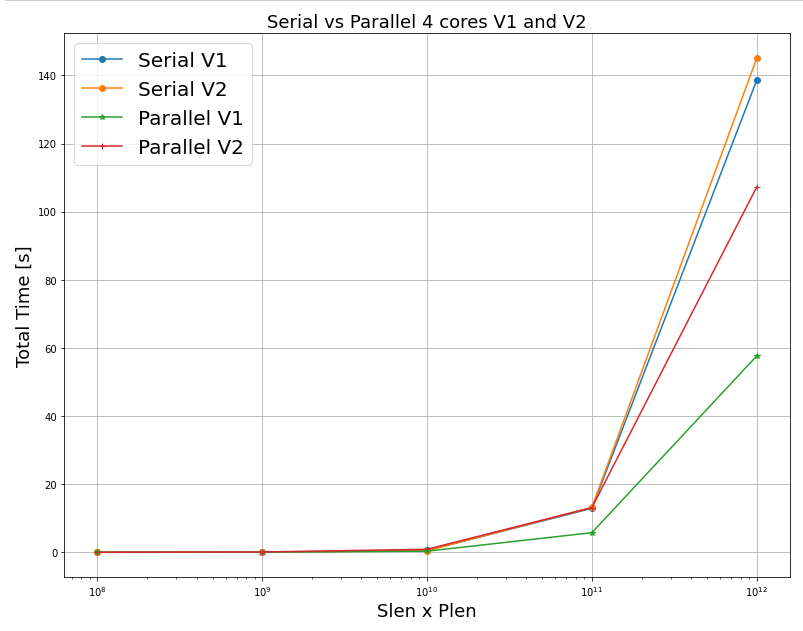
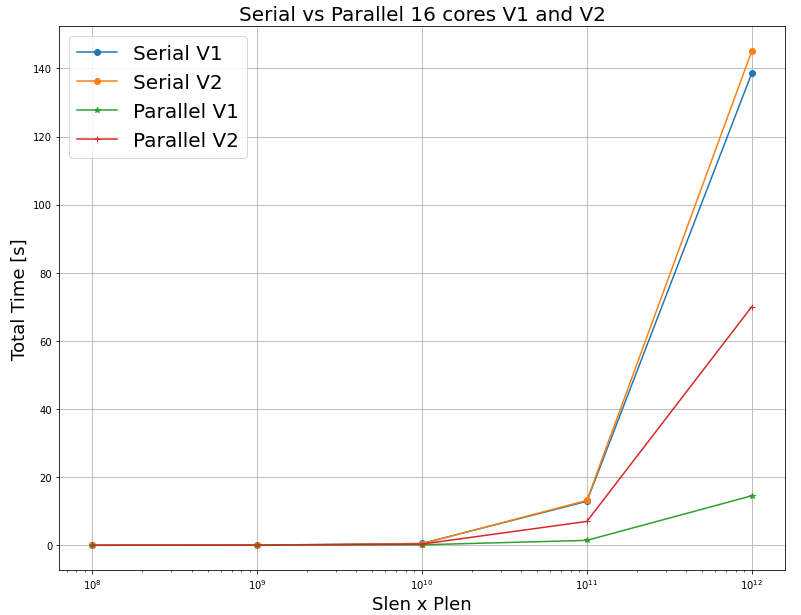
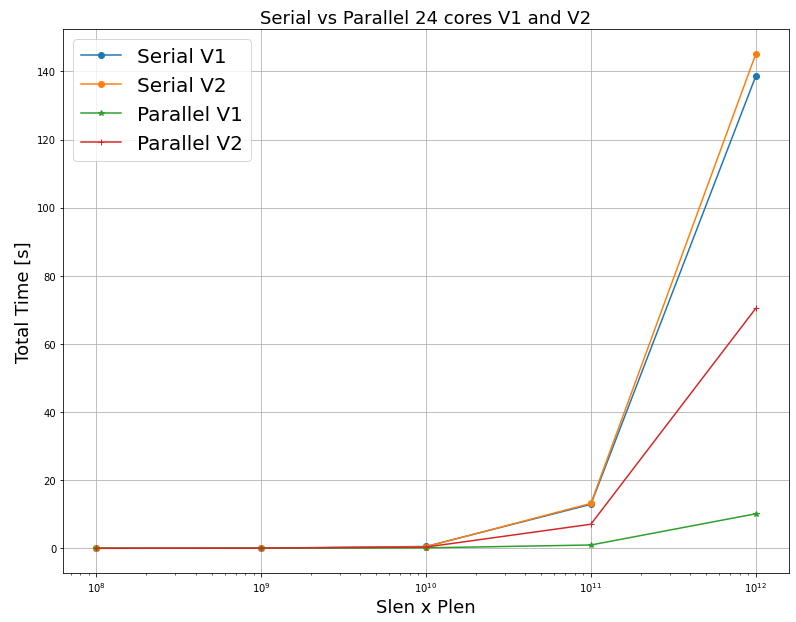
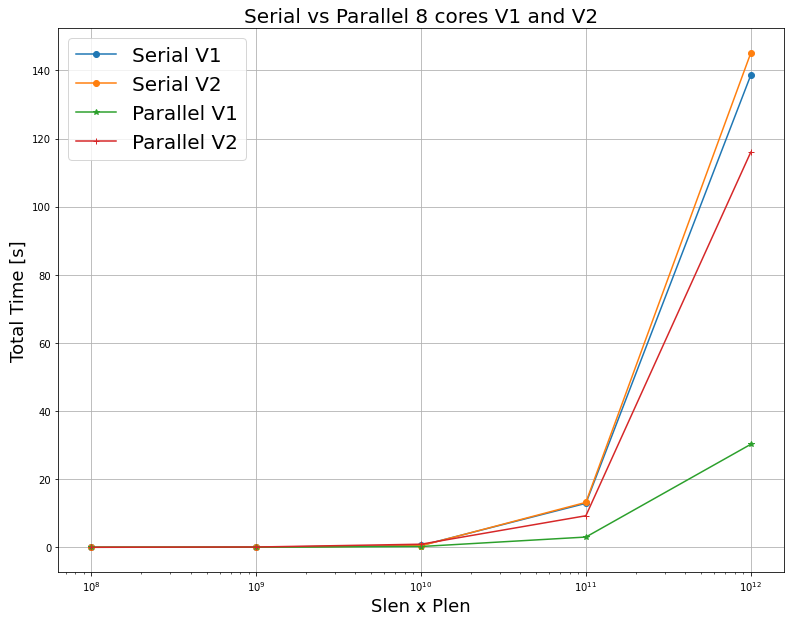
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Core | Match time V1 | Total time V1 | Expected Speedup | Real S |
|  | 2 | 0.00394 | 0.00436 | 1.34515 | 0.0871 |
| S 10^4 | 4 | 0.00044 | 0.00074 | 1.62570 | 0.5135 |
| P 10^2 | 8 | 0.00154 | 0.00190 | 1.81497 | 0.2 |
|  | 16 | 0.00377 | 0.00399 | 1.92715 | 0.09523 |
|  | 24 | 0.01158 | 0.01175 | 1.96769 | 0.03234 |
|  | 2 | 0.00379 | 0.00515 | 1.62614 | 1.1281 |
| S 10^5 | 4 | 0.00224 | 0.00371 | 2.36725 | 1.5660 |
| P 10^3 | 8 | 0.01183 | 0.01338 | 3.06589 | 0.4342 |
|  | 16 | 0.00160 | 0.00288 | 3.59662 | 2.0173 |
|  | 24 | 0.00374 | 0.00490 | 3.81686 | 1.1857 |
|  | 2 | 0.03619 | 0.04642 | 1.68163 | 1.09543 |
| S 10^6 | 4 | 0.02125 | 0.03421 | 2.55108 | 1.4864 |
| P 10^3 | 8 | 0.01153 | 0.02391 | 3.4405 | 2.1267 |
|  | 16 | 0.02231 | 0.03225 | 4.16687 | 1.5767 |
|  | 24 | 0.00440 | 0.01441 | 4.48231 | 3.5287 |
|  | 2 | 0.37222 | 0.46656 | 1.7069 | 1.1497 |
| S 10^7 | 4 | 0.20925 | 0.32978 | 2.64002 | 1.6266 |
| P 10^3 | 8 | 0.10835 | 0.23258 | 3.63307 | 2.3065 |
|  | 16 | 0.05404 | 0.14792 | 4.47465 | 3.6266 |
|  | 24 | 0.03672 | 0.13258 | 4.84907 | 4.0462 |
|  | 2 | 10.15439 | 10.24702 | 1.98577 | 1.2598 |
| S 10^7 | 4 | 5.64013 | 5.75996 | 3.91583 | 2.2413 |
| P 10^4 | 8 | 2.87570 | 3.01167 | 7.61792 | 4.2866 |
|  | 16 | 1.35811 | 1.45170 | 14.44728 | 8.8929 |
|  | 24 | 0.88490 | 0.97405 | 20.60449 | 13.2538 |
|  | 2 | 101.74417 | 102.76884 | 1.98699 | 1.3486 |
| S 10^8 | 4 | 56.45102 | 57.65937 | 3.92292 | 2.4037 |
| P 10^4 | 8 | 29.04837 | 30.23910 | 7.64931 | 4.5833 |
|  | 16 | 13.60532 | 14.53676 | 14.56874 | 9.5132 |
|  | 24 | 9.20338 | 10.13168 | 20.858 | 13.6795 |

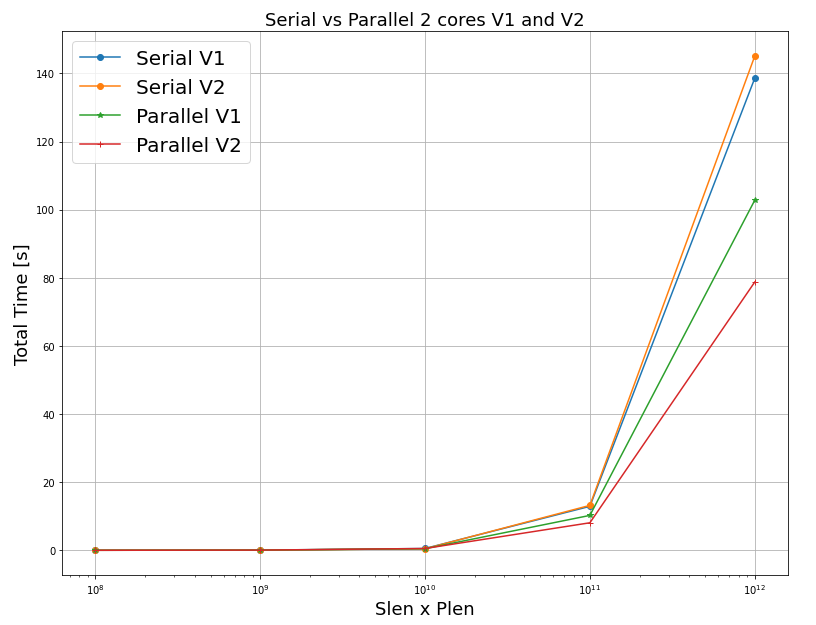
Table 2: Parallel version V1 - analysis result.

Just for precision, the .

As it was expected, by using more vCPUs we obtain much better performances (clearly visible in the last few rows of the table).

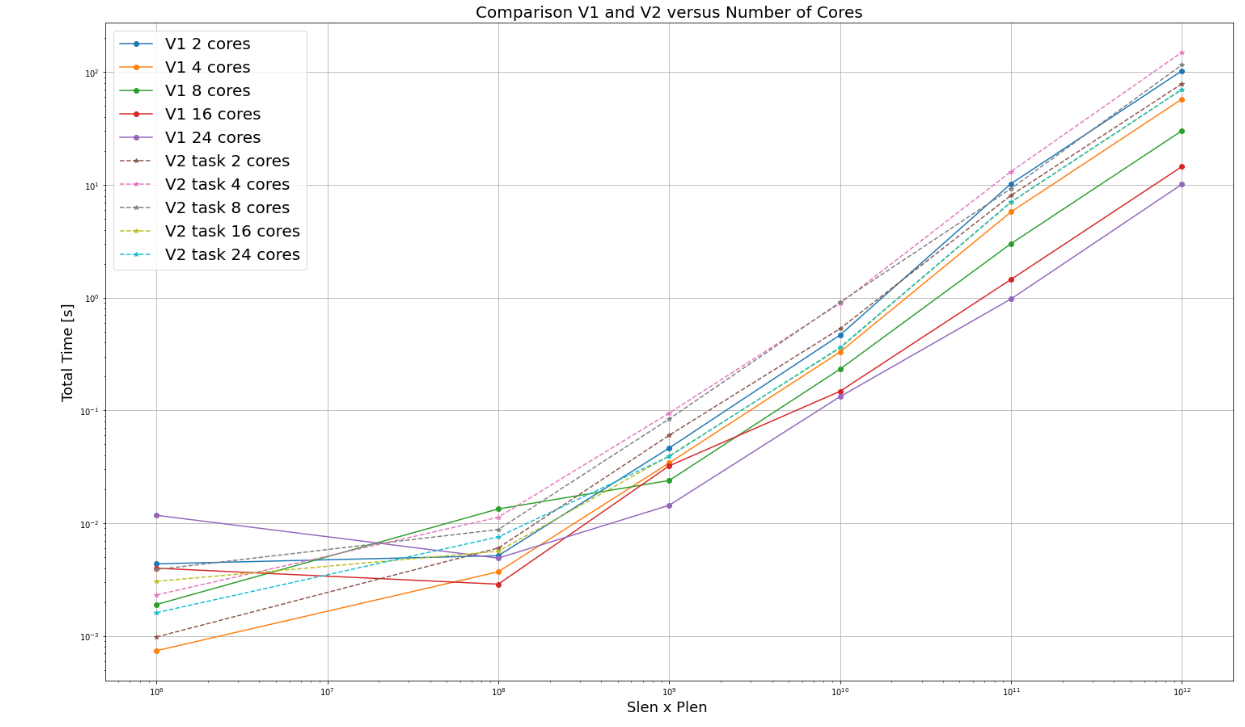
A relevant aspect is that the theoretical speedup tends to the number of cores, and thus we have a potentially highly parallelizable code which is a good feature for our program. Anyways, there exists a difference between the expected speedup calculated by the Amdahl’s law and the effective one. This is most probably due to synchronization of threads and overhead in their management, that is not taken into account by the theoretical formula. As for the second parallel version with tasks, the results obtained are much worse than expected. We still have an improvement with respect to both the serial versions, but the first parallel version V1 behaves better in almost all the cases, as can be seen in the Figures 2-6. In these plots we show a comparison of the two parallel versions against the serial codes, while changing the number of cores to 2, 4, 8, 16 and 24 respectively.



The only situation in which parallel version V2 is able to outperform version 1 is when using 2 cores.

As a summary of all these considerations we can analyse Figure 7, below. Once again, we can notice that the dashed lines (representing version v2) yield on average worse performances than those represented by continuous lines (i.e.: version v1).

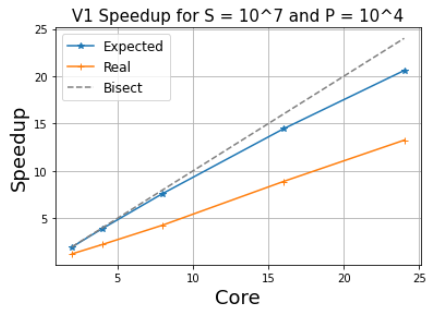
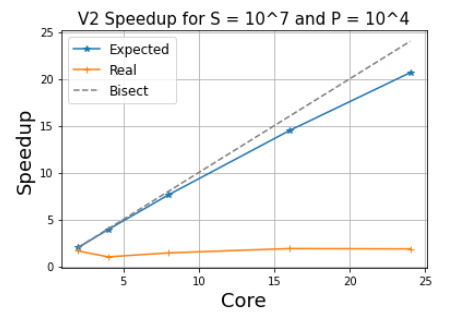
However, selecting smaller *‘slen’* and *‘plen’* the behaviour is more unstable as we have worse timings for a larger number of cores. In fact, our problem is a quite simple one and a better performance is much more visible in higher sizes of the sequences.



A final remark has still to be made on the speedup. In the following Table 3, we report the behaviour of the real speedup compared to what we theoretically expected in both versions v1 and v2 when using more cores. We can evaluate the speedup for the same example we analysed before: 16 vCPUs, slen=10^7, plen=10^4. Considering the fastest serial code as a reference, the expected value is 14.5. We already know the real improvement will be slightly less than this value and in particular coincides with 8.9 for version v1, which is still a satisfactory result.  
In the second parallel version, instead, what we get as real speedup is a value equal to 1.88 that is clearly a poor improvement.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 2 vCPUs | 4 vCPUs | 8 vCPUs | 16 vCPUs | 24 vCPUs |
| Expected S Serial V1 | 1.98577 | 3.91583 | 7.61792 | 14.44728 | 20.60449 |
| Real S V1 Par | 1.2598 | 2.2413 | 4.2866 | 8.8929 | 13.2538 |
| Real S V2 Par | 1.63323 | 1.000546 | 1.42374 | 1.88296 | 1.85 |

Holding fix slen equal to 10^7 and plen equal to 10^4, the Speedups obtained are the following (compared with the Serial V1 that is pretty equal to the serial V2)

Figures 8 and 9 (below), show graphically the speedup in function of the number of cores. Both for V1 and V2 the growing of the speedups calculated with the Amdahl’s’ Law are linear, while the Real Speedup are sub-linear. From this representation is clear another time, why we prefer the solution 1.