# New High Performance Solutions for the Multiple Genetic Sequence Alignment Problem

Mario João Jr.*+, Vinod Rebello* (PhD Advisor), Alexandre Senna+ (Mentor)
*Fluminense Federal University, +State University of Rio de Janeiro

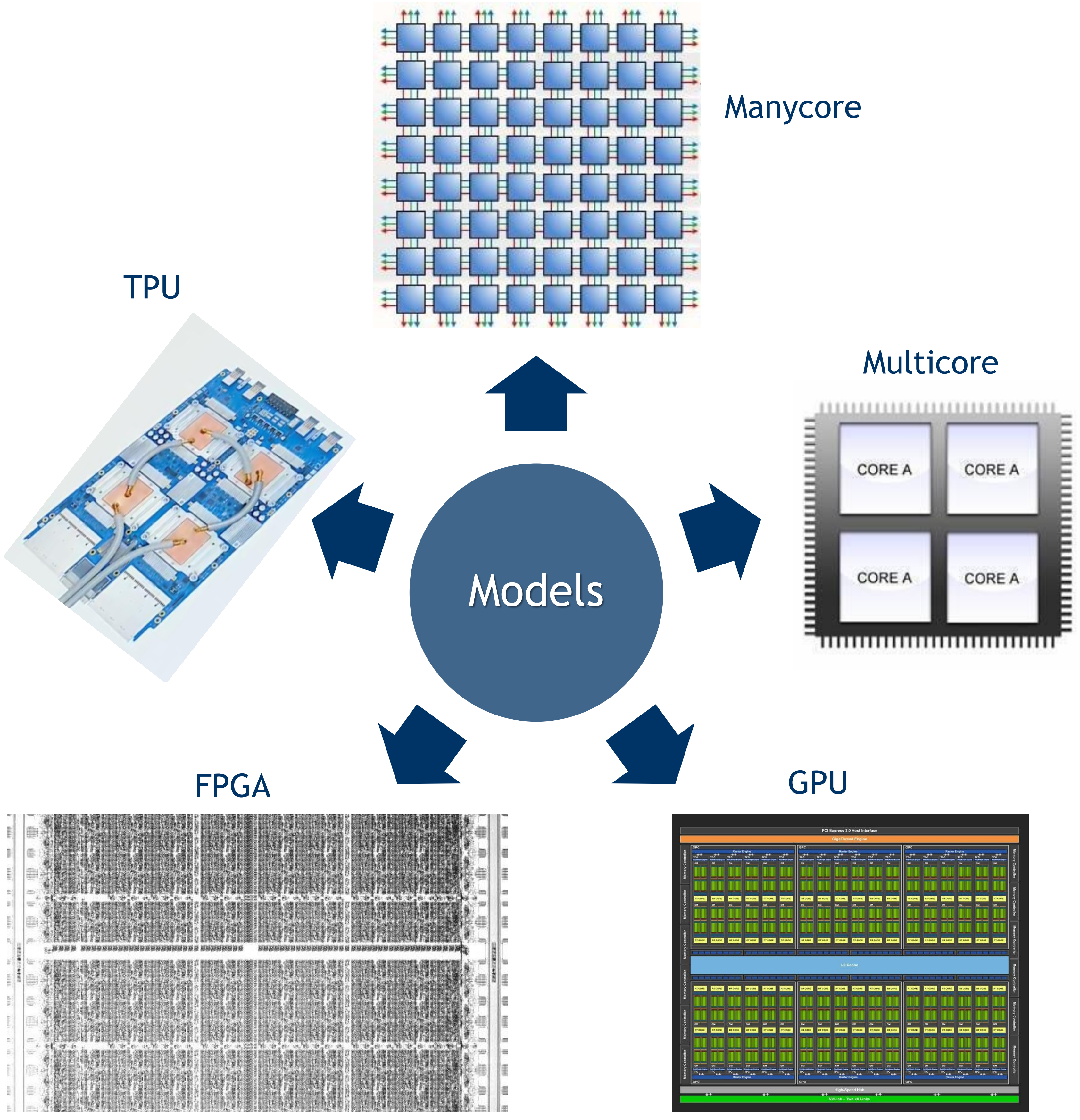
## Problem Description

- Multiple genetic sequence alignment (MSA) is the alignment of three or more biological sequences
- It is a NP-complete problem
- Dynamic Programming algorithms for best alignment have complexity O($l^n$) where n is the number of sequences to be aligned and l is the mean length of the sequences
- It is used in areas of computational biology such as: phylogenetics, evolutionary analysis, biodiversity, structural and functional inference of proteins and homology searches
- Example:

```
GAA-ATATTTTCATTTCAGAGGTAGTCAGT---------GAGGATCCATCATCACCACGAGAGG
GAAGGTGTGAACCATCCCTATGCCAGCCGAAAGTTGCGCCACCATCTCGCCTCTATAACAAGGG
GAACGAATGGACCCGACTGAAGACGGCGCACCGCACCGCGAGGGTACGGGCACGACAACGATGG
----------ACCTAGCTGCTGGCATCAGATA-------CTGCATTCAGATCCAACAAGACTGG
```

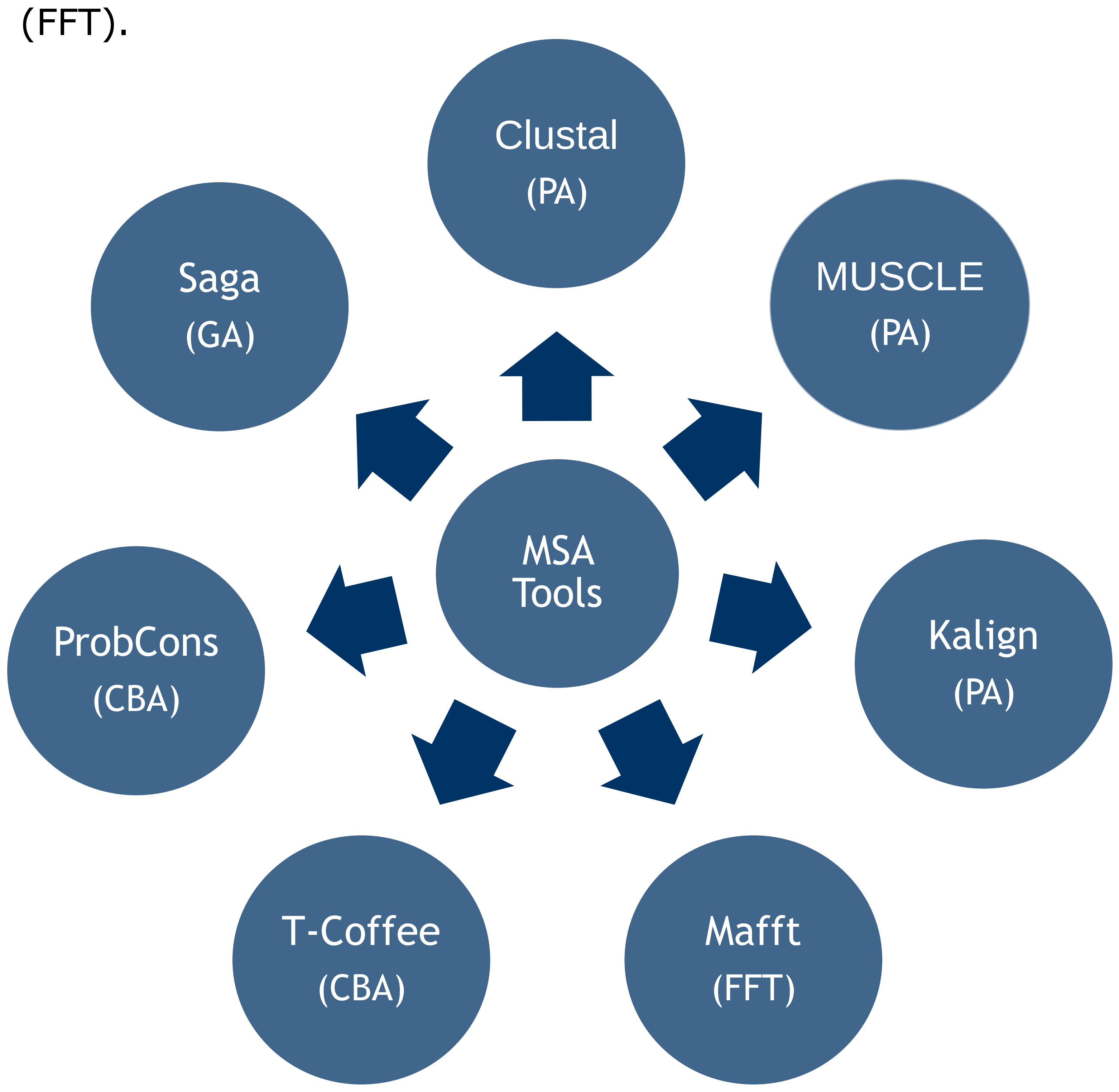
## Research Goals

- Propose and evaluate a novel MSA algorithm that obtains accurate results, in computationally reasonable times, for the alignment of hundreds of sequences of sizes that can reach thousands of bases, genes or proteins;
- Parallelization techniques must be identified that exploit the characteristics of parallel computing architecture models available both now and in the near future;
- Actual techniques include Wave front, Bag of Tasks, Loop Optimizations and use of intrinsic instructions;
- Our initial steps will focus on the analysis of existing tools and their strategy to exploit parallelism;
- Then, our first solution will attempt to handle pair-to-pair scores generation for the manycore architectural model.
- Some of architectural models that currently exist include:



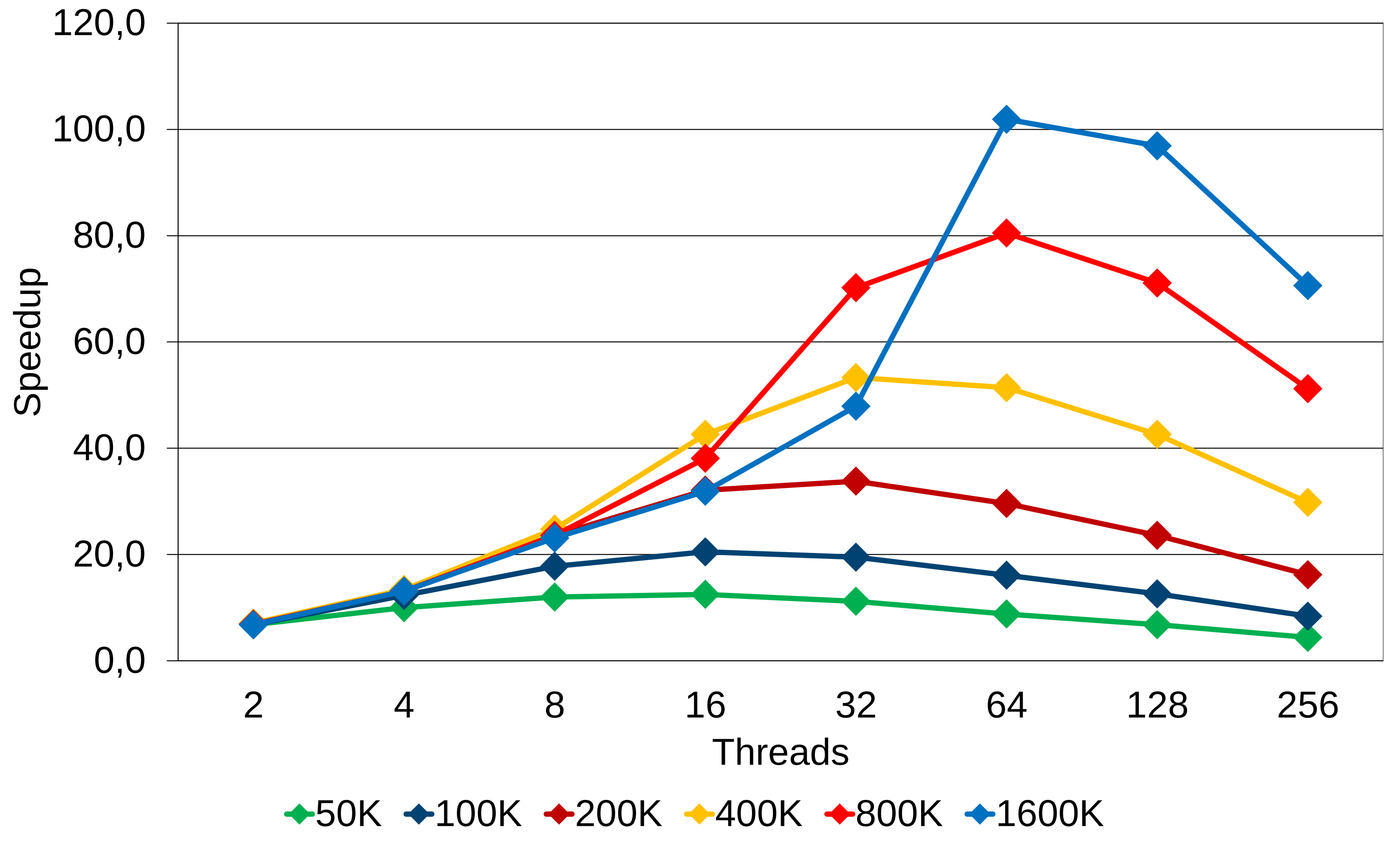
## Well Known Tools and their Heuristics

- Current tools are predominantly based on heuristics such as Progressive Alignment (PA), Consistence-Based Alignment (CBA), Genetic Algorithms (GA) or a Fast Fourier Transform (FFT).



## Previous Experience

- Our research work [1] has so far has focused on the alignment of two sequences and improving the implementation of the Hirschberg Algorithm for Longest Common Subsequence (LCS) problem.
- We have proposed three parallelization approaches to better adapt its execution on Manycore architectures and exploit SIMD vector instruction sets.
- After analyzing and redesigning the algorithm, a speedup of 105, compared to the original sequential version, was reached using 68 processing cores and sequences with 1.6 million characters.



## Reference

[1] M. João Jr., A. C. Sena, and V. E. F. Rebello, "On the parallelization of Hirschberg's algorithm for multi-core and manycore systems," Concurrency and Computation: Practice and Experience, Early View, doi:10.1002/cpe.5174

Universidade Federal Fluminense
Instituto de Computação
SGC LAB SMART GRID COMPUTING LAB
UNIVERSIDADE DO ESTADO DO RIO DE JANEIRO UERJ
DTIES Departamento de Tecnologias da Informação e Ensino em Saúde L@MPADA