Overview
Suffix array construction is one of the most fundamental full-text indices with many applications in text indexing, data compression and bioinformatics. The problem boils down to sorting all suffixes of a given text in lexicographic order. Although closely related to the more general problem of string sorting, the suffix sorting problem can be solved with work only linear in the text length by exploiting the suffixes’ internal structure.

The problem has received a lot of attention, especially in the sequential and shared-memory parallel setting. However, with the ever increasing amount of data, distributed-memory algorithms – which are able to efficiently use tens of thousands of processors – become more and more important.

While there exist linear work distributed-memory suffix array construction algorithm [4] (at least for specialized settings), they have not been implemented in a scaling fashion and the only practical implementations are quasilinear algorithms [3] [2] [1]. The challenge of developing fast distributed suffix sorters lies in designing and engineering algorithms that efficiently distribute the workload and minimize communication overheads among the processors.

Objective
The main objective of this master’s thesis is to design, implement and engineer new scalable distributed suffix sorting algorithms building on our previous work on distributed suffix and string sorting [2] [5].

Requirements
- Good C++ and MPI programming skills
- Interest in string and distributed algorithms

References