Abstract: The goal of this thesis is to propose a layer on top of edge-cloud, in order to provide soft real-time guarantees on the execution time of applications. This is done in order to satisfy the soft-real time requirements set by the developers of latency-sensitive applications. The proposed layer uses a predictor of execution time, in order to find combinations of processes with, which satisfy the soft real-time requirements when collocated. To implement the predictor, we are provided with information about the resource usage of processes and execution times of collocated combinations. We utilize similarity between the processes, cluster analysis, and regression analysis to form four prediction methods. We also provide a boundary system of resource usage used to filter out combinations exceeding the capacity of a computer. Because the metrics indicating the resource usage of a process can vary in their usefulness, we also added a system of weights, which estimates the importance of each metric. We experimentally analyze the accuracy of each prediction method, the influence of the boundary detection system, and the effects of weights.