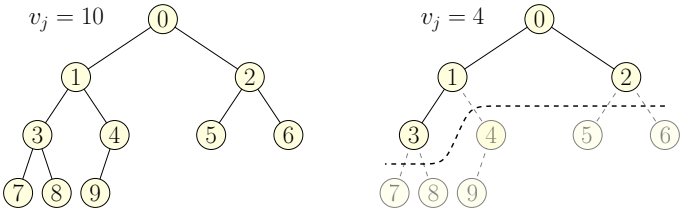
K-Means in a Malleable Distributed Environment (DE/EN)



Our **decentralized job scheduling** platform *Mallob* enables to resolve NP-hard problems "on demand" in a large distributed environment (thousands of cores). A key technology used by Mallob is the ability of each job to support a fluctuating number of parallel workers during its processing (*malleability*). Each job j is organized as a binary tree T_j of workers which grows and shrinks dynamically according to the job's current *fair volume* v_j (see above illustration). So far, Mallob features a highly competitive engine for propositional satisfiability (SAT solving). In order to investigate the impact of our model of malleability for different applications, we want to integrate engines for further problems into our system.

K-Means is a popular clustering method and, as such, a crucial building block for many machine learning methods. Given n d-dimensional points and a parameter k, we want to cluster the points into k groups (clusters) while minimizing the sum of squared distances from each point to its cluster center. Finding an optimal solution to this problem is NP-hard. The K-Means algorithm initializes cluster centers randomly and then alternates between assigning each point to the nearest cluster and updating each cluster's center to the center of weight of its members.

We offer a **bachelor or master thesis** which covers the following tasks:

- Design a distributed variant of the K-Means algorithm which can handle a varying number of parallel workers during its execution.
- Implement the algorithm as a new application engine in the Mallob system.
- Evaluate your algorithm experimentally and analyze how malleable scheduling of resources impacts its performance.

Solid programming skills in C++ are required – it may be possible to acquire these skills over the course of the project. Basic knowledge on MPI is beneficial.

A thesis can be supervised in German or English and should be written in English. If interested, please contact Dominik Schreiber ⟨dominik.schreiber@kit.edu⟩.

References

dominikschreiber.de/papers/2021-sat-scalable.pdf
dominikschreiber.de/papers/2022-submission-europar-mallob.pdf
en.wikipedia.org/wiki/K-means_clustering