

# PARALLELIZATION IN MULTIPLE IMPUTATION

Sven Nekula, Joshua Simon and Eva Wolf

Otto Friedrich University, Bamberg



## What is Parallelization?

**Parallelization** is a technique to fasten time-consuming computations. It uses several cores on a **CPU** parallelly and splits up the computational work on them. Afterwards, the results are merged. This can reduce the runtime significantly. Parallelization is nevertheless limited to a certain kind of tasks: The computations to be parallelized must be independent from each other, as information cannot be cross-accessed over the cores while the process is running.

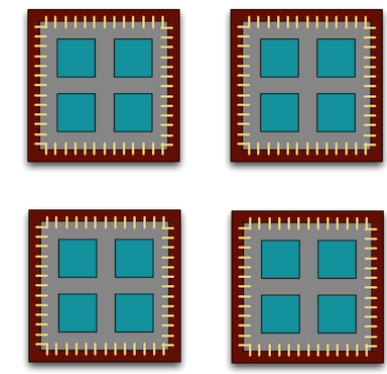


Figure 1: 4 processors with 4 cores each

## Theory

In **parallel programming**, the multiple cores of a computational system want to be used best to decrease computation time. Gene **Amdahl** was the first one to describe the boundaries of that project: Every parallel process also requires additional workload, so called "data management housekeeping". The speed up through parallel processing will tend to 0 at a certain amount of processing units involved, as this **overhead** workload exceeds the capacity of the computational unit which it is assigned to.

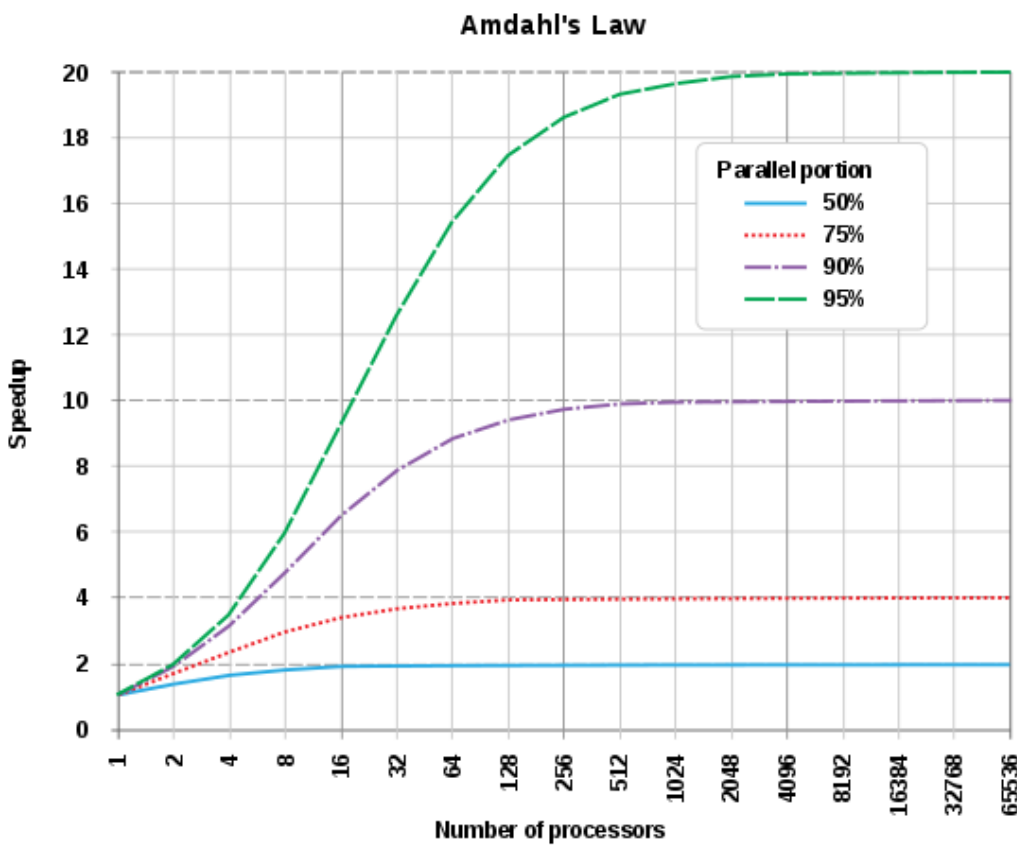


Figure 2: Amdahls Law

**Multiple imputation** is a method to complete a dataset with missing information. It relies on the estimation of the missing values through different methods. What is common to them is that we use not a single imputation run, but several. The results of all imputation runs are then merged and lead to realistic uncertainty of the estimators given the missing data. It is prone to be parallelized, as the several imputation runs can be processed independent of each other on different cores and easily merged afterwards.

## Methodology

As the **dataset** we used a simple data generator of normally distributed random variables. The data set created contains 10 variables, of which some depend on each other. The sample size is  $n=10000$ . Parallelization favors complex data sets, so the amount of variables and their interdependency should not be too small.

**Time measurement** was done with the `system.time` function, which returns 3 values: User CPU-, User System- and Elapsed time. User CPU is the time needed for the current task such as an execution in R. System CPU describes the time needed by the operating system to organize that task such as opening folders or asking for the System time. Elapsed time is the wall clock time that passed while the function and background processes were running.

The **method** of the **mice** algorithm was set on default, pmm.

The **speed up** value is calculated by the serial time (runtime without parallelization) divided by the runtime of the current parallelization implementation [cite:chapple2016mastering]. Values below one show that the parallelized run took longer than the serial run.

## Results

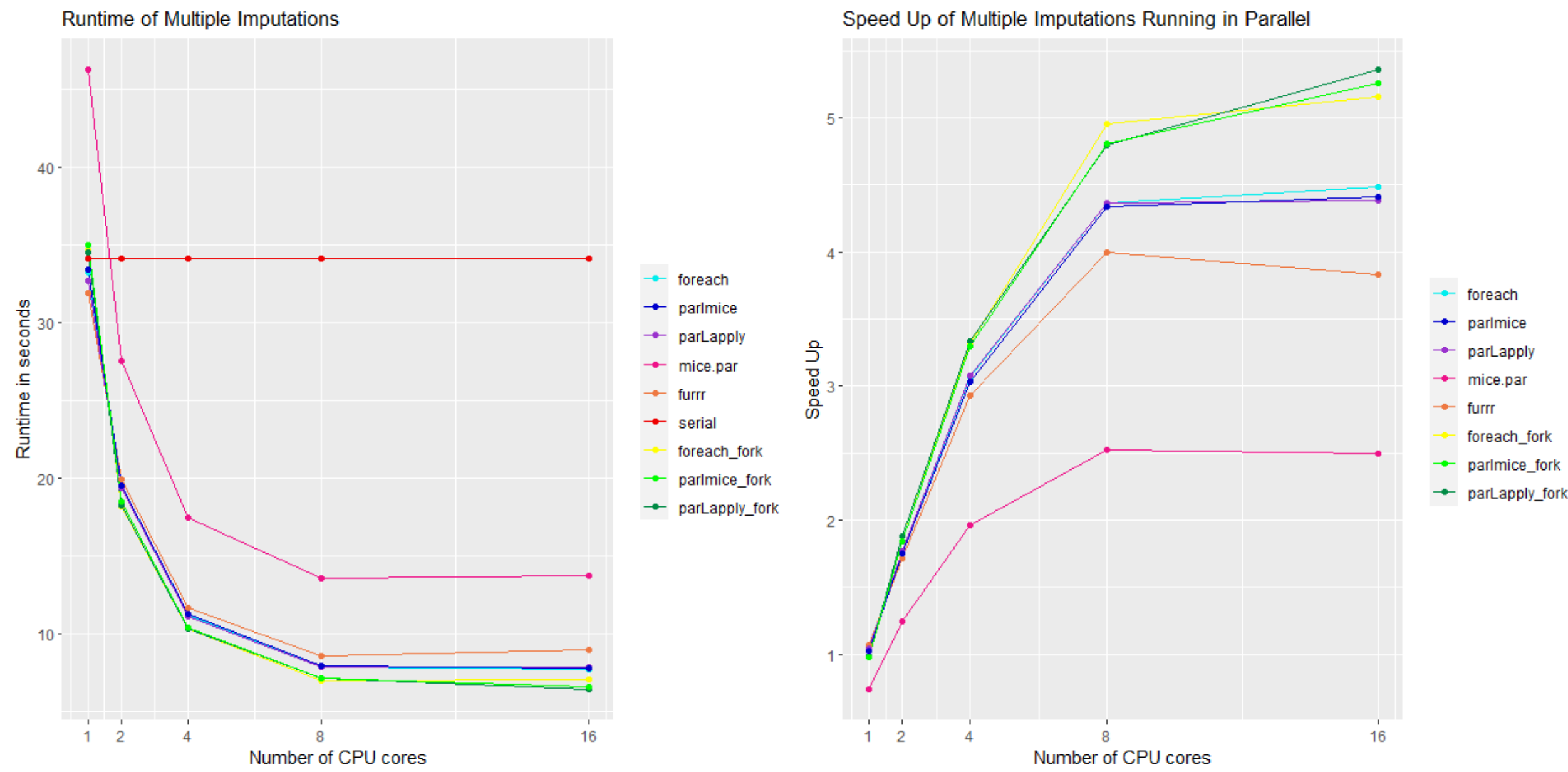


Fig. 1: Runtime and Speedup from 1 up to 8 cores

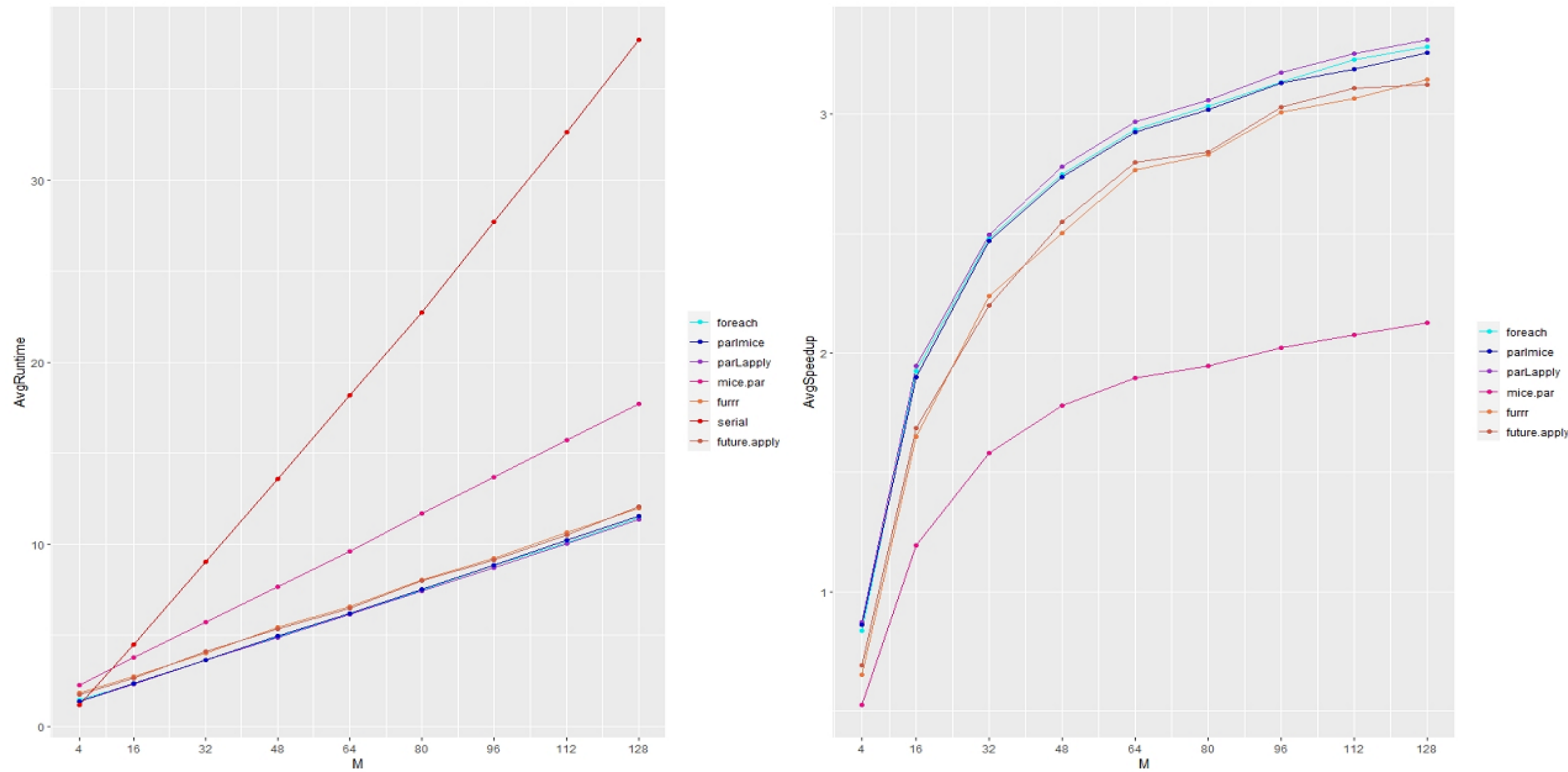


Fig. 2: Runtime and Speedup from 1 up to 128 Imputation runs

Function	Cores	Avg. user time	Avg. system time	Avg. elapsed time	Avg. speed up
1 foreach	16	0.25	0.03	7.63	4.48
2 foreach_fork	16	0.21	0.17	7.03	5.15
3 furr	16	1.07	0.09	8.93	3.83
4 mice.par	16	1.15	0.10	13.67	2.50
5 parLapply	16	0.33	0.04	7.80	4.39
6 parLapply_fork	16	0.23	0.17	6.40	5.36
7 parlmice	16	0.30	0.02	7.76	4.41
8 parlmice_fork	16	0.21	0.18	6.54	5.25
9 serial	16	34.11	0.03	34.13	1.00

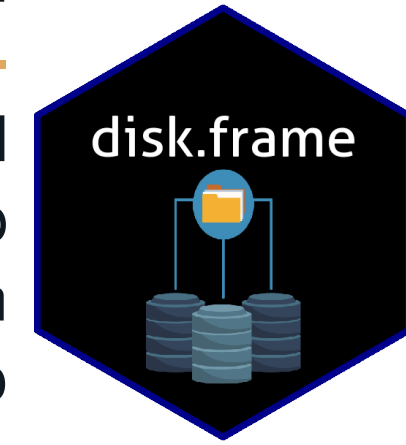
Tab. 1: Time components for the different implementations on 16 cores

## Comparison of Implementations

**foreach::foreach** usability  
**mice::parlmice** usability  
**micemd::mice.par** usability  
**parallel::parLapply** usability  
**furrr::future\_map** usability  
**future.apply::future\_lapply** usability  
*PSOCK vs. FORK*

## Further Applications

The opposite case to parallelized processes is the intended time-intensive, serial processing of Big Data with **disk framing**. In that setting we have data that exceeds the **RAM capacity** of the system, so that the data needs to be split up into several smaller chunks. These are then processed in portions that fit the available memory. Nevertheless, also for **disk framing**, parallelization is a useful tool to decrease computation time. The system chooses as many chunks of data as they fit into the RAM and then processes them parallelly, so that use of the computational power of the system can be made, while the memory capacity limits the speed by the amount of data which is processed at once.



Considering today's trends in parallel and high performance computing, **scalability** plays a big role. So-called distributed systems are used here. These can be seen as the fusion of several computers into one system, often referred to as a **computing cluster**. In order to run a computer program on such an architecture, a **message passing interface (MPI)** is used, which now handles the communication between multiple CPUs and their cores. Popular applications are cloud solutions like Amazon's AWS (Amazon Web Services). The R language can make use of MPI with packages like **Rmpi** or **pbdsMPI** [cite:chapple2016mastering] and together with the **OpenMPI** implementation utilize the lower level of parallelism [cite:robey2021parallel].

## References