Welcome to the course

PARALLEL PROGRAMMING IN R

Hana Sevcikova

Senior Research Scientist, University of Washington

Prerequisites

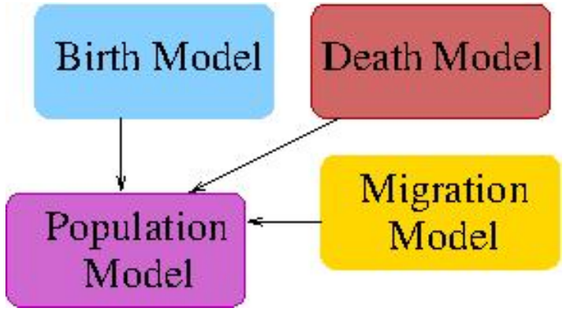
- Writing Efficient R code
- Optimized sequential code
- Benchmark your code

Overview

1. Methods of parallel programming & supporting R packages
2. The `parallel` core package in detail
3. Packages `foreach` and `future.apply`
4. Random numbers & reproducibility and final example

Splitting computation problems for parallel processing

I. By Tasks



II. By Data

$$\begin{bmatrix} 1 & 8 & 13 & 12 \\ 14 & 11 & 2 & 7 \\ 4 & 5 & 16 & 9 \\ 15 & 10 & 3 & 6 \end{bmatrix}$$

Summary of partitioning

1. **By Task:** Apply different tasks to the same or different data.
2. **By Data:** The same task is performed on different data.

Example (splitting by data):

$$1 + 2 + 3 + ... + 100$$

```
sum(1:25) + sum(26:50) + sum(51:75) + sum(76:100)
```

Embarrassingly parallel applications

$$\text{Many such independent tasks} = \text{embarassingly parallel}$$

E.g., many statistical simulations of the structure (in pseudo-code):

```
initialize.rng()
for (it in 1:N) result[it] <- myfunc(...)
process(result, ...)
```

Let's practice!

PARALLEL PROGRAMMING IN R

Models of parallel computing

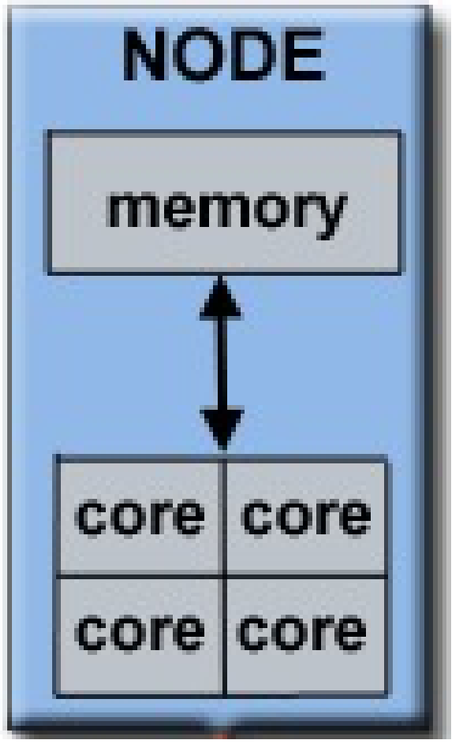
PARALLEL PROGRAMMING IN R

Hana Sevcikova
University of Washington

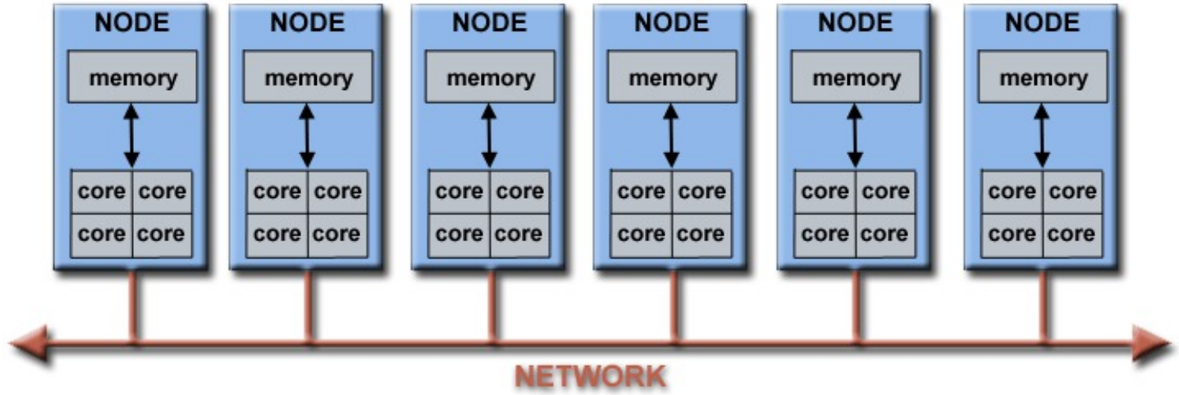
datacamp

Hardware - Central processing unit (CPU)

I. **Multi-processor** (CPU, core) computer

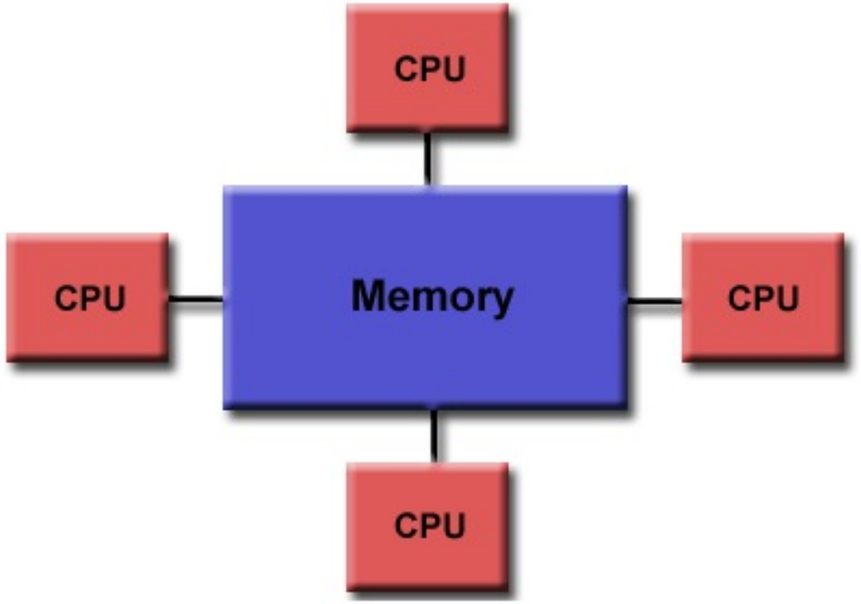


II. **Cluster** of single- or multi-processors computers



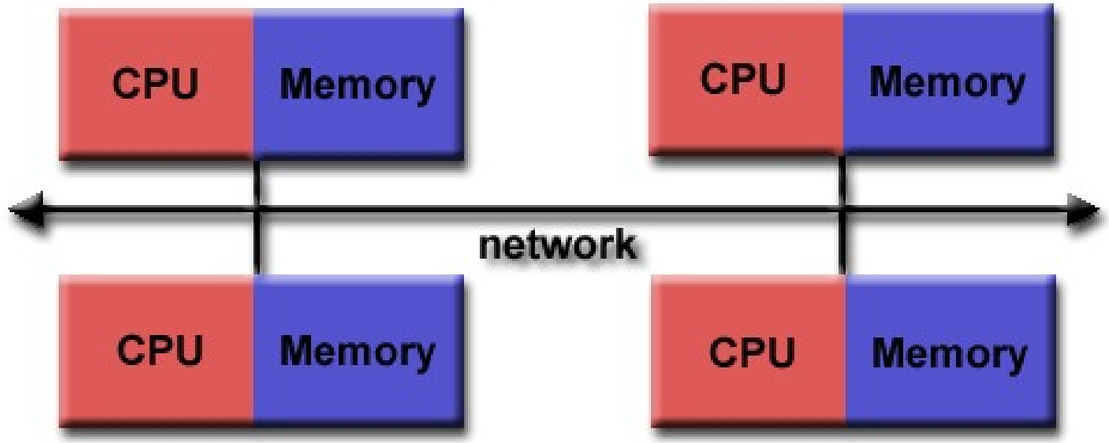
Hardware - Memory

- **Shared memory**



  - Shared memory software
  - Message-passing software

- **Distributed memory**



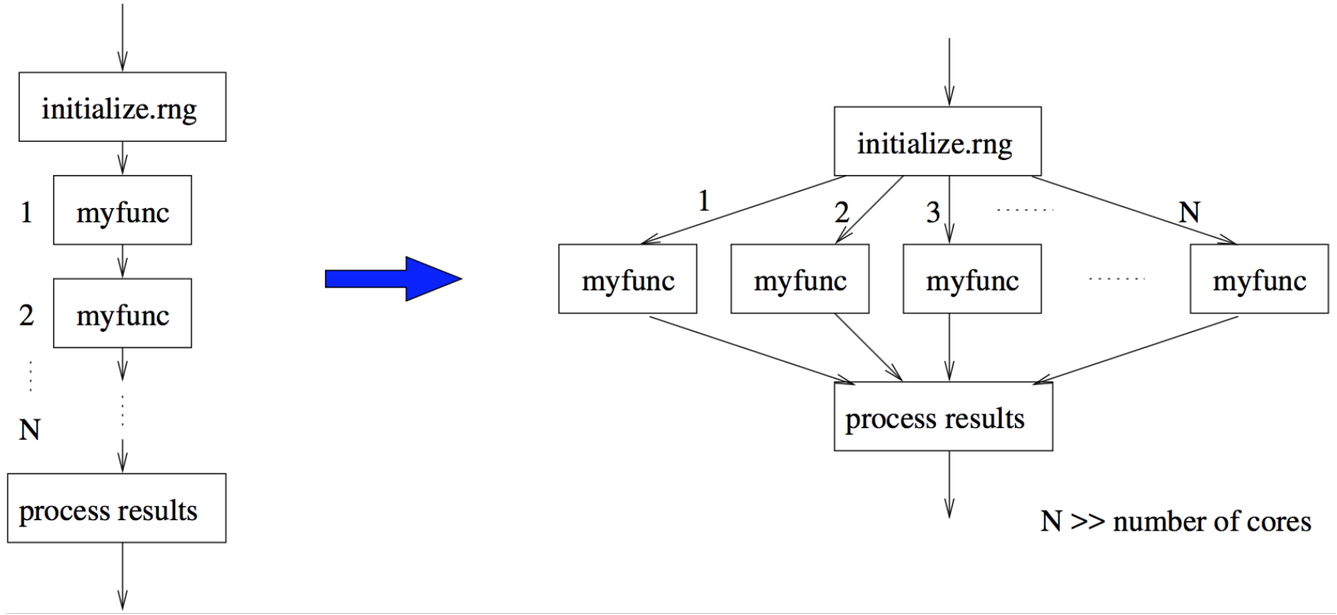
  - Message-passing software

Programming paradigms

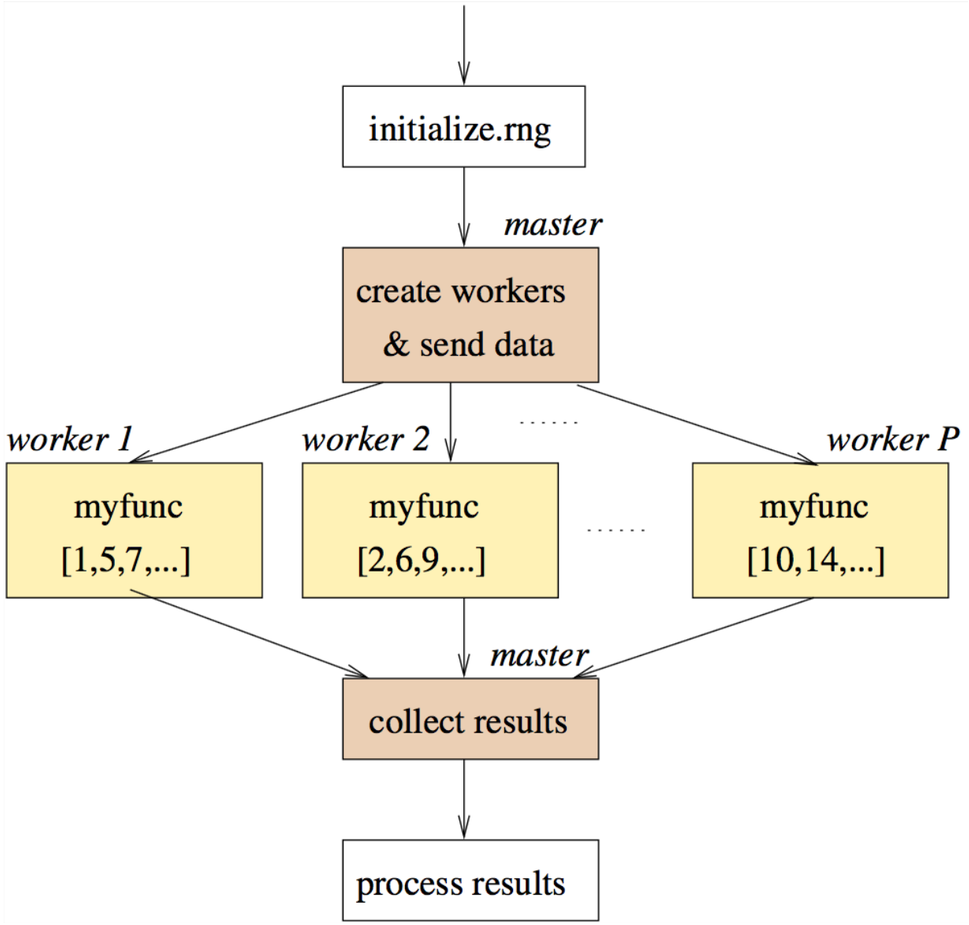
- Master-worker model
- Map-reduce paradigm
  - applications for distributed data
  - Hadoop, Spark
  - **Scalable Data Processing in R**

Master-worker model (1)

```
initialize.rng()
for (it in 1:N) result[it] <- myfunc(...)
process(result, ...)
```



Master-worker model (2)



Let's practice!

PARALLEL PROGRAMMING IN R

R packages for parallel computing

PARALLEL PROGRAMMING IN R

Hana Sevcikova
University of Washington

datacamp

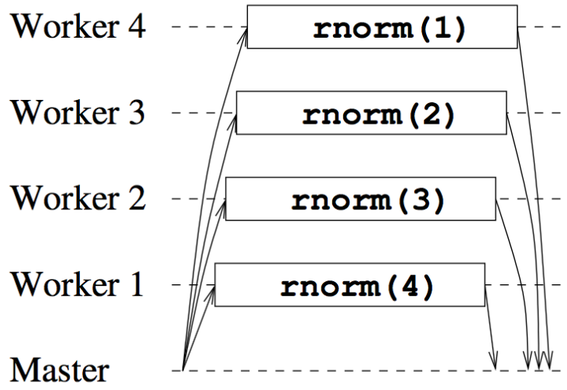
R packages

- Core package: `parallel`
- Parallel support for big data:
  - `sparklyr` , `iotools`
  - `pbdR`
- Embarrassingly parallel, master-worker model:
  - `foreach` , `future.apply`
  - `snow` , `snowFT` , `snowfall`
  - `future`

Package parallel

```
library(parallel)
ncores <- detectCores(logical = FALSE)
cl <- makeCluster(ncores)
clusterApply(cl, x = ncores:1, fun = rnorm)
stopCluster(cl)
```

ncores = 4 → x = c(4, 3, 2, 1)



Let's practice!

PARALLEL PROGRAMMING IN R