A gentle introduction to parallel computing in R

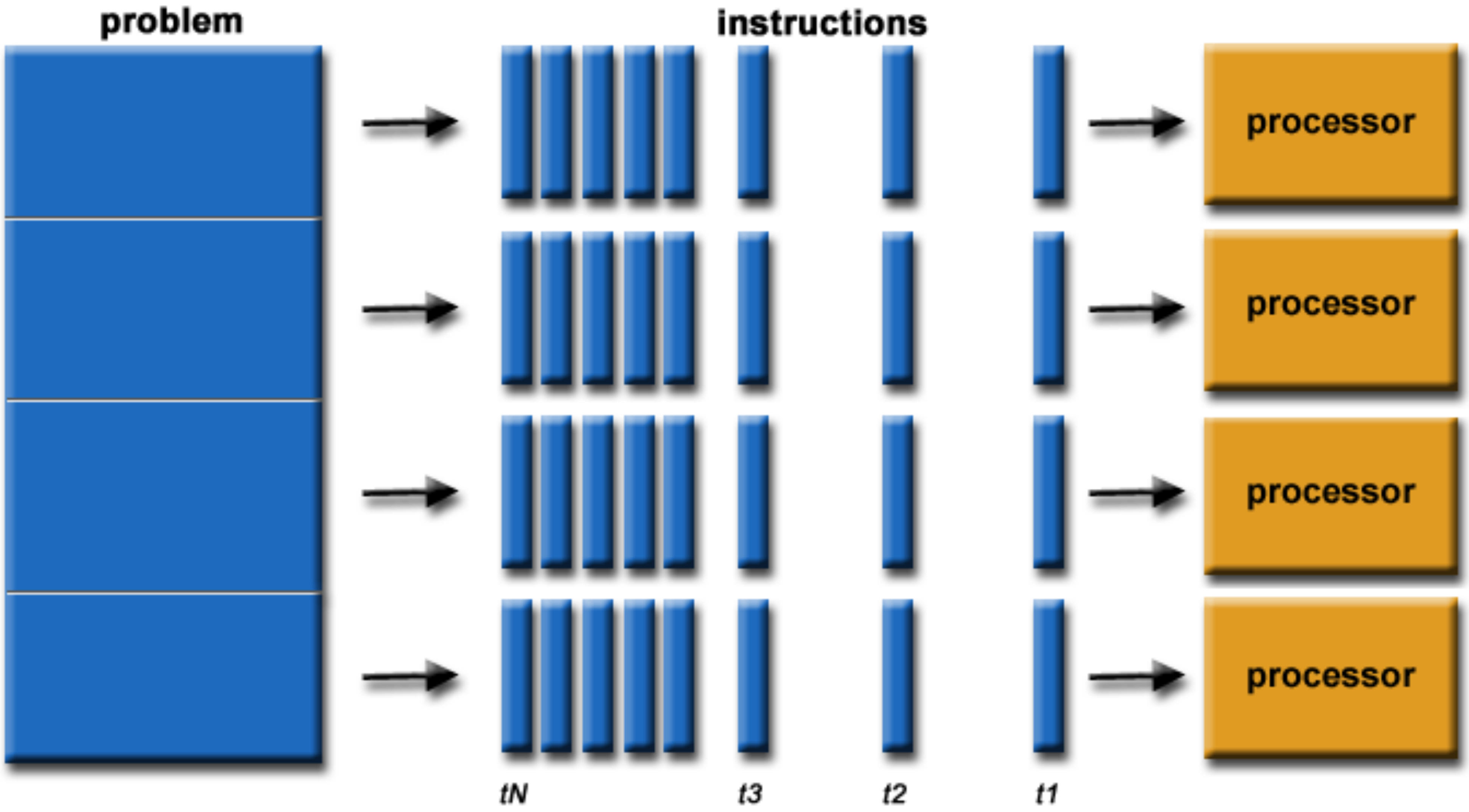
REC Annual Meeting

October 25, 2018

Malcolm Itter

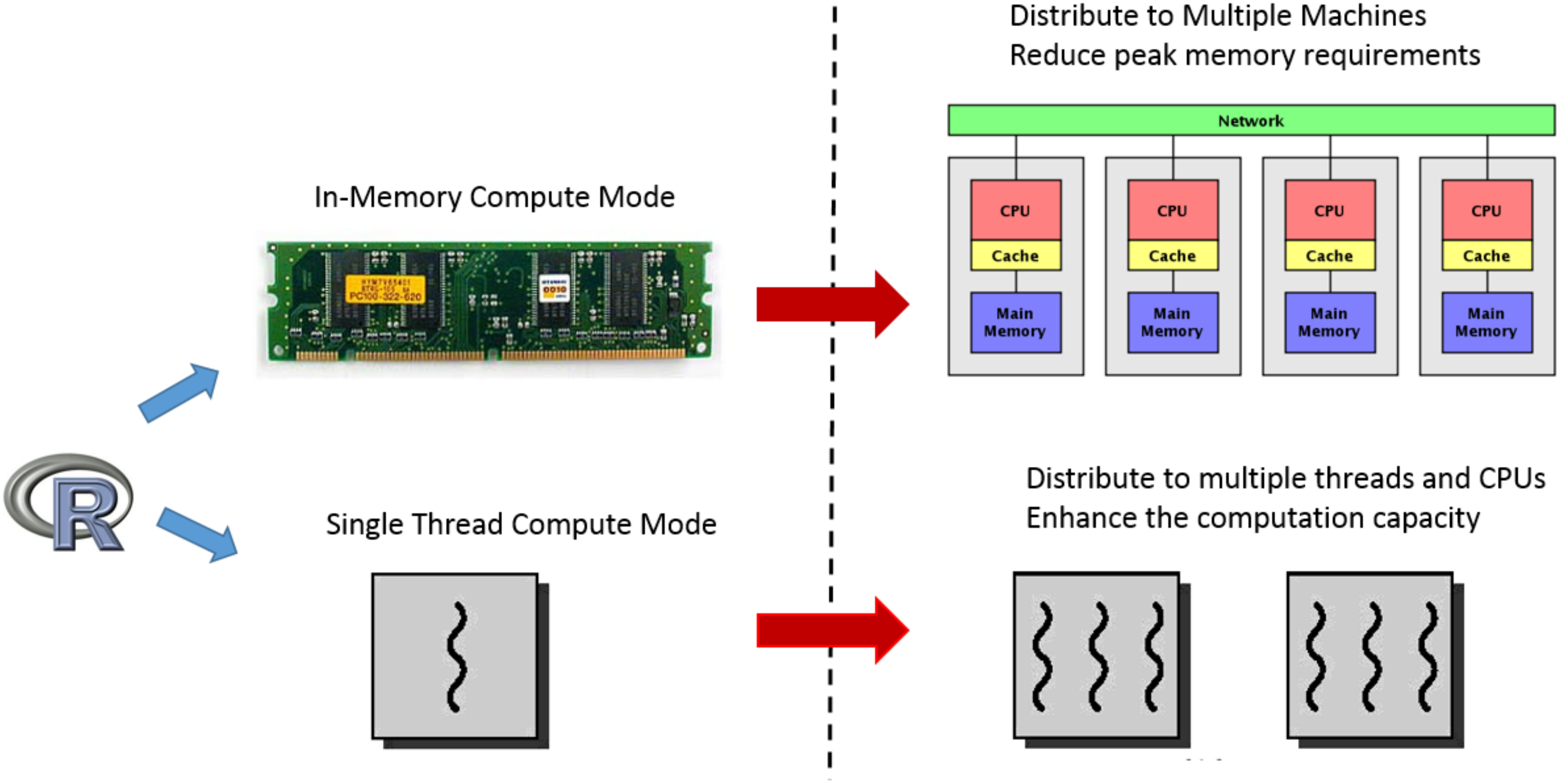
What is parallel computing

Break process into discrete parts, each with its own instructions, complete parts in parallel…



Credit: Blaise Barney, Lawrence Livermore National Laboratory

Parallel computing in R



In-Memory Compute Mode

Distribute to Multiple Machines
Reduce peak memory requirements

Single Thread Compute Mode

Distribute to multiple threads and CPUs
Enhance the computation capacity

R core

Parallel Computing Technology

Credit: Peng Zhao, R with Parallel Computing from User Perspectives

Objectives

- Provide basic overview of parallel computing tools in R
- Designed to get user up-and-running with parallel operations
- Focus on applications, not underlying computing structure (e.g., parallel vs. distributed system) or theory
- Computing tasks throughout are motivated by improving ecological inference
- By end, you should be able to run your R code in parallel lowering the bar to use computationally-intensive methods

What we will cover

Generally,

1. Embarrassingly parallel processes —> problems that are easily broken into discrete parts

2. Focus on shared memory multiprocessing systems: single computer with memory that may be simultaneously accessed by one or more programs running on multiple CPUs (e.g., your laptop or desktop!)

3. We'll let R functions take care of the shared memory, creation of master/slave processes, and communication among CPUs

What we will cover

Specifically,

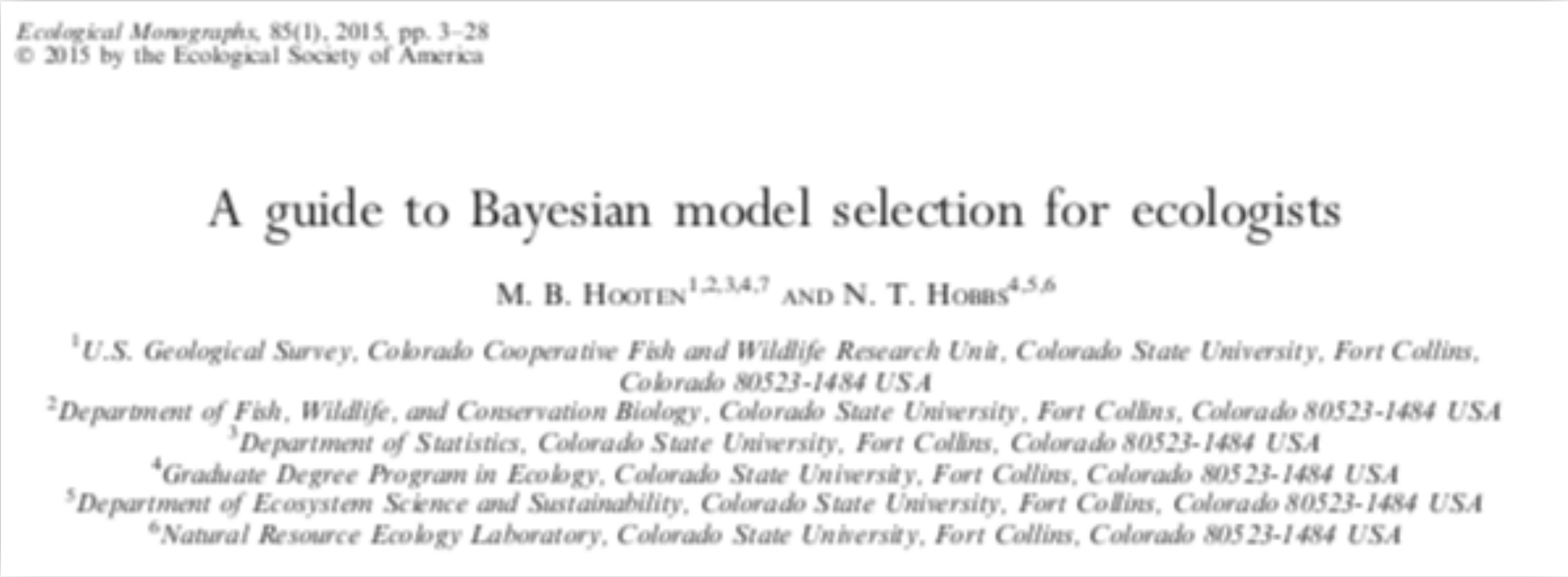
1. Vectorized vs. non-vectorized operations
2. **`apply`** family of functions (e.g., **`apply`**, **`lapply`**, **`sapply`**)
3. Tools to benchmark R code (**`sys.time`**, **`tictoc`**, **`rbenchmark`**)
4. Parallel computing using **`snow`**/**`snowfall`**

Motivating analysis

- Throughout, we will use data on willow tit occurrence in the Swiss alps as a motivating analysis
- We will utilize parallel computing within R to help us:
  - Fit a Bayesian hierarchical site-occupancy model
  - Apply the model selection techniques described in Hooten and Hobbs (2015)

Willow tit (*Parus montanus*)

Image: Wiki Commons



Ecological Monographs, 85(1), 2015, pp. 3–28
© 2015 by the Ecological Society of America

A guide to Bayesian model selection for ecologists

M. B. Hooten[1,2,3,4,7] and N. T. Hobbs[4,5,6]

[1] *U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, Colorado 80523-1484 USA*
[2] *Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, Colorado 80523-1484 USA*
[3] *Department of Statistics, Colorado State University, Fort Collins, Colorado 80523-1484 USA*
[4] *Graduate Degree Program in Ecology, Colorado State University, Fort Collins, Colorado 80523-1484 USA*
[5] *Department of Ecosystem Science and Sustainability, Colorado State University, Fort Collins, Colorado 80523-1484 USA*
[6] *Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, Colorado 80523-1484 USA*

Useful references/resources

- Rossini et al. (2007) Simple parallel statistical computing in R. *Jour. of Comp. and Graph. Stat.* **16(2)**: 399-420. https://doi.org/10.1198/106186007X178979
- Knaus, J. (2010) Developing parallel programs using snowfall. *R vignette*. https://CRAN.R-project.org/package=snowfall
- Paciorek, C. Tutorial on parallel & distributed computing. https://github.com/berkeley-scf/tutorial-parallel-distributed
- Hammerling, D. and Finley, A. (2018) High performance computing for spatial data. http://blue.for.msu.edu/envr18/
- Dirk Eddelbuettel (creator of Rcpp) Webpage: http://dirk.eddelbuettel.com/presentations/