What to do when a computing problem is too big, or computing would take too long, for a single computer to handle

- parallel computing
- using graphical processing units for mathematical calculations
- distributed and grid computing

Parallel computing in R: the snow package

- snow is a parallel computing package for R
- snow uses a master/worker model:
  - The user starts an ordinary R session as the master process.
  - This session creates (or connects to) a set of worker processes.
  - Jobs are sent to the worker processes and results are returned.
- The underlying message passing can be based on
  - raw sockets (no additional packages/software needed)
  - PVM (uses rpvm and PVM)
  - MPI (uses Rmpi and LAM-MPI; other MPIs may also work)
- Which communication mechanism is used only matters at startup.


Simple example of parallel computing in R

```
# R session on node 11 of the statistics cluster

> library(snow)

# establish communication with 10 other cores on 3 nodes

> cl <- makeSOCKcluster( c("node11","node11","node11","node12","node12","node12","node13","node13","node13"))

# execute the Sys.info() function on all 10 worker cores, and retrieve just the nodename from it

> do.call("rbind", clusterCall(cl, function(cl) Sys.info()"nodename"))

node11.beowulf.stat.uiowa.edu
node11.beowulf.stat.uiowa.edu
node11.beowulf.stat.uiowa.edu
node12.beowulf.stat.uiowa.edu
node12.beowulf.stat.uiowa.edu
node12.beowulf.stat.uiowa.edu
node13.beowulf.stat.uiowa.edu
node13.beowulf.stat.uiowa.edu
node13.beowulf.stat.uiowa.edu
node13.beowulf.stat.uiowa.edu

# time a computation on this vector done sequentially on the master node

> x <- 1:100/101

# system.time(qtukey(x, 2, df=2))

user  system  elapsed
```

General programming with graphical processing units

- high end graphics cards such as those used in gaming computers have many cores for rendering graphics (up to several hundred)
- recently languages and programming toolkits have been developed to enable the use of the cores in graphics cards for mathematical computation
- challenges are the same as parallel processing on a cluster
  - has to be possible to divide the whole computing job (or parts of it) into individual small tasks that can be executed independently in parallel
  - slow transfer of data and results between CPU (regular processor) and GPU (graphics processor)
- stay tuned for some examples from my project group

Distributed or grid computing

- distributed system: system of autonomous computers that communicate through a computer network to work together to solve a common computational problem
- distributed computing: use of distributed systems to solve computational problems
  - whole problem split into individual tasks, each carried out by single computer
- organizational grids
- volunteer computing

The TeraGrid: example of a scientific grid

See some slides in separate lecture.
Volunteer computing

- distributed computing in which computer owners donate some of their computer resources to one or more projects
- what kinds of computing problems are suitable (Kate)
- BOINC software (Yawa)
- World Community Grid
  - an example of a well-run volunteer grid computing project
  - purpose and organization

Scientific projects currently active on World Community grid

- Help Cure Muscular Dystrophy - Phase 2
- Help Fight Childhood Cancer
- Clean Energy Project - Phase 2
- Computing for Clean Water
- Help Conquer Cancer
- Human Proteome Folding - Phase 2
- FightAIDS@Home
- Discovering Dengue Drugs - Together - Phase 2

Discovering Dengue Drugs Together – Phase 2

- family of viruses called flaviviruses causes Dengue, Hepatitis C, West Nile, Yellow Fever, and other diseases
- caring for people infected by these diseases is a strain on world’s health care
  - no effective drugs to treat these diseases
  - dengue, West Nile, and yellow fever transmitted by different species of mosquitoes
  - 40% of world’s population is at risk for dengue
  - each year about 1.5 million people are treated for dengue, dengue hemorrhagic fever
  - about 2% of world’s people are infected with hepatitis C

- mission of DDDT2 is to identify promising drug candidates to combat flaviviruses
  - enzyme called protease is essential for virus replication
  - uses structure-based drug discovery calculations to identify chemicals that may inhibit protease in flaviviruses
- computing can be used to screen huge databases of chemicals
- promising ones found can then go through laboratory testing and possibly proceed to clinical testing