

Nagiza F. Samatova, Marcia Branstetter, Auroop R. Ganguly, Robert Hettich, Shiraj Khan, Guruprasad Kora, Jiangtian Li, Xiaosong Ma, Chongle Pan, Arie Shoshani, and Srikanth Yeginath

## Motivation behind Parallel R

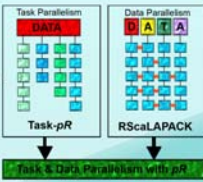
- Ideal Programming Requirements:**
  - Be able to use existing high level (i.e. R) code
  - Require minimal extra efforts for parallelizing
  - Have identical/similar (presumably easy-to-use) interface to R's
  - Be able to test codes in sequential settings
  - Provide efficient and scalable (in terms of problem size and number of processors) performance

### Task-parallel analyses:

- Likelihood Maximization.
- Re-sampling schemes: Bootstrap, Jackknife, etc.
- Animations
- Markov Chain Monte Carlo (MCMC).
- Multiple chains.
- Simulated Tempering: running parallel chains at different "temperature" to improve mixing.

### Data-parallel analyses:

- k-means clustering
- Principal Component Analysis (PCA)
- Hierarchical (model-based) clustering
- Distance matrix, histogram, etc. computations



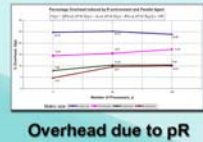
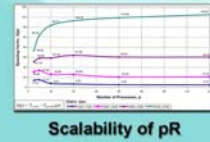
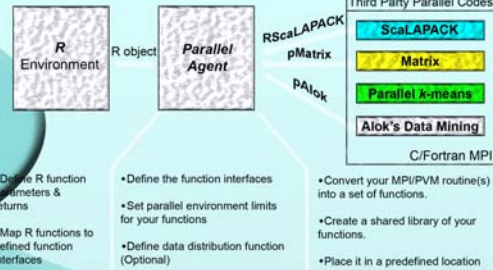
```

fileList<-list.files(pattern="*.nc");
for (i in 1:length(fileList)) {
  matrix [i] <- readNcFile (fileList[i]);
  pca [i] <- prcomp (matrix [i])
}
R

fileList<-list.files(pattern="*.nc");
PE ( for (i in 1:length(fileList)) {
  matrix [i] <- readNcFile (fileList[i]);
  pca [i] <- sla.prcmp (matrix [i])
}
pR
  
```

Data size, n	h	h(h-1)	n <sup>2</sup>
100B	10 <sup>10</sup> ops	10 <sup>10</sup> ops	10 <sup>10</sup> ops
100K	10 <sup>10</sup> ops	10 <sup>10</sup> ops	10 <sup>10</sup> ops
100M	10 <sup>10</sup> ops	10 <sup>10</sup> ops	10 <sup>10</sup> ops
100B	10 <sup>10</sup> ops	10 <sup>10</sup> ops	10 <sup>10</sup> ops
100M	10 <sup>10</sup> ops	10 <sup>10</sup> ops	10 <sup>10</sup> ops
100B	10 <sup>10</sup> ops	10 <sup>10</sup> ops	10 <sup>10</sup> ops
100M	10 <sup>10</sup> ops	10 <sup>10</sup> ops	10 <sup>10</sup> ops

## Extensibility of Parallel R



## Geo-statistical and Spatial Data Analysis with GRASS and pR

Use Case: G. Fann, J. Drake, B. Budhend

- Leverages the work by Markus Neteler ([http://grass.itc.it/statsgrass/grass\\_geostats.html](http://grass.itc.it/statsgrass/grass_geostats.html))
- Offers a richer set of statistical analysis capabilities including (Basic Statistics, Exploratory Data Analysis, Linear Models, Multivariate Analysis, Time Series Analysis, etc.)
- Provides high performance and parallel computational platform for large datasets

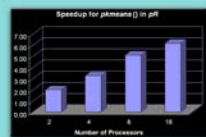
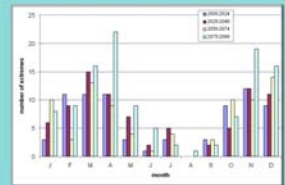
## Quantitative Proteomics in Biology with pR

Use Case: C. Pan and R. Hettich

### Genes and pathways involved in aromatic degradation

## Climate Extremes Analysis with pR

Use Case: A. Ganguly, M. Branstetter, S. Khan



Contact: Nagiza Samatova, samatovan@ornl.gov

<http://www.aspect-sdm.org/Parallel-R/>