Introducing a New Client/Server Framework for Big Data Analytics with the R Language

Drew Schmidt, Wei-Chen, and George Ostrouchov

July 19, 2016

Acknowledgements

- Support
  - This work was supported in part by the project "Harnessing Scalable Libraries for Statistical Computing on Modern Architectures and Bringing Statistics to Large Scale Computing" funded by the National Science Foundation Division of Mathematical Sciences under Grant No. 1418195.
  - This work used resources supported by the National Science Foundation under Grant Number 1137097 and by the University of Tennessee through the Beacon Project.
  - Just got an allocation on Comet! Thank you XSEDE!

- Disclaimer
  - The findings and conclusions in this presentation have not been formally disseminated by the U.S. Department of Health & Human Services nor by the U.S. Department of Energy, and should not be construed to represent any determination or policy of University, Agency, Administration and National Laboratory.

1 Background and Motivation

2 pbdR Compute Backend

3 The Client/Server

4 Challenges and Future Work

Problems with "Big Data" Software

- *Many* frameworks; what do they all do?
- Don't always play nice with HPC systems.
- Often not as "high level" as advertised.
- Almost exclusively batch!

Data Analysis Is An Interactive Activity

Data analysis is an interactive activity[a]

[a] Data analysis is an interactive activity

Why am I here?

- XSEDE: a conference for users and service providers.
- I want to talk to you!

1 Background and Motivation

2 pbdR Compute Backend

3 The Client/Server

4 Challenges and Future Work

pbdR

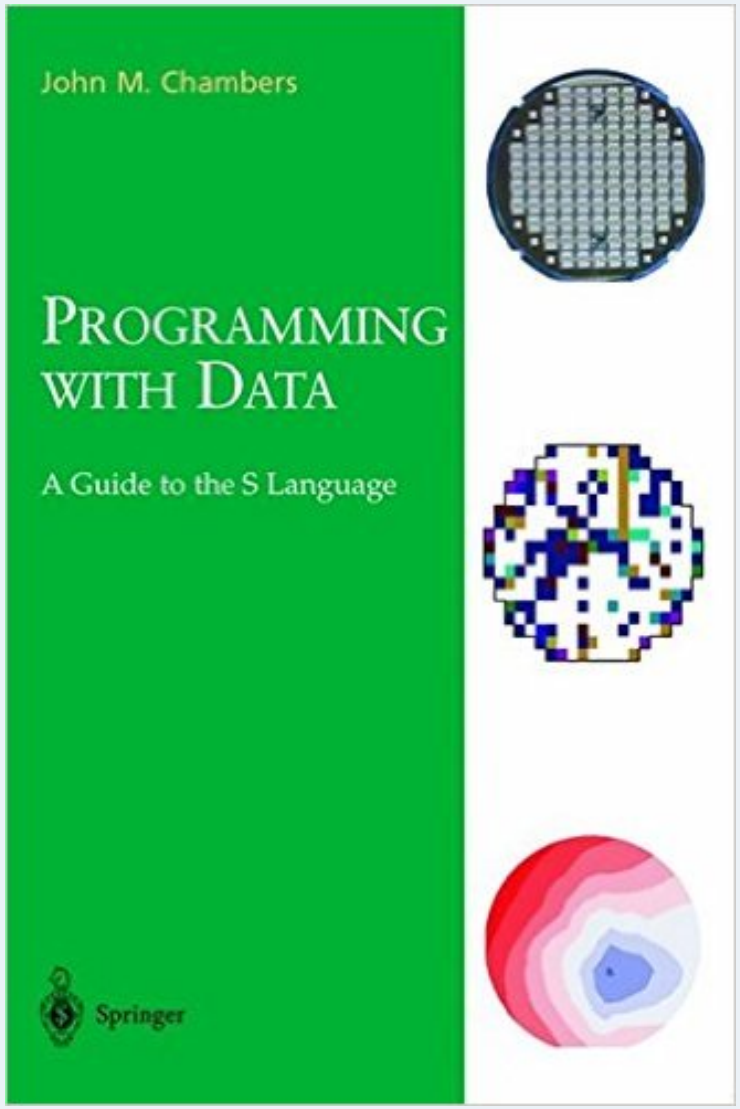
Client/Server + **pbdR** = interactive big data analysis

- This talk is not about the old stuff...
- But we have to talk about it!

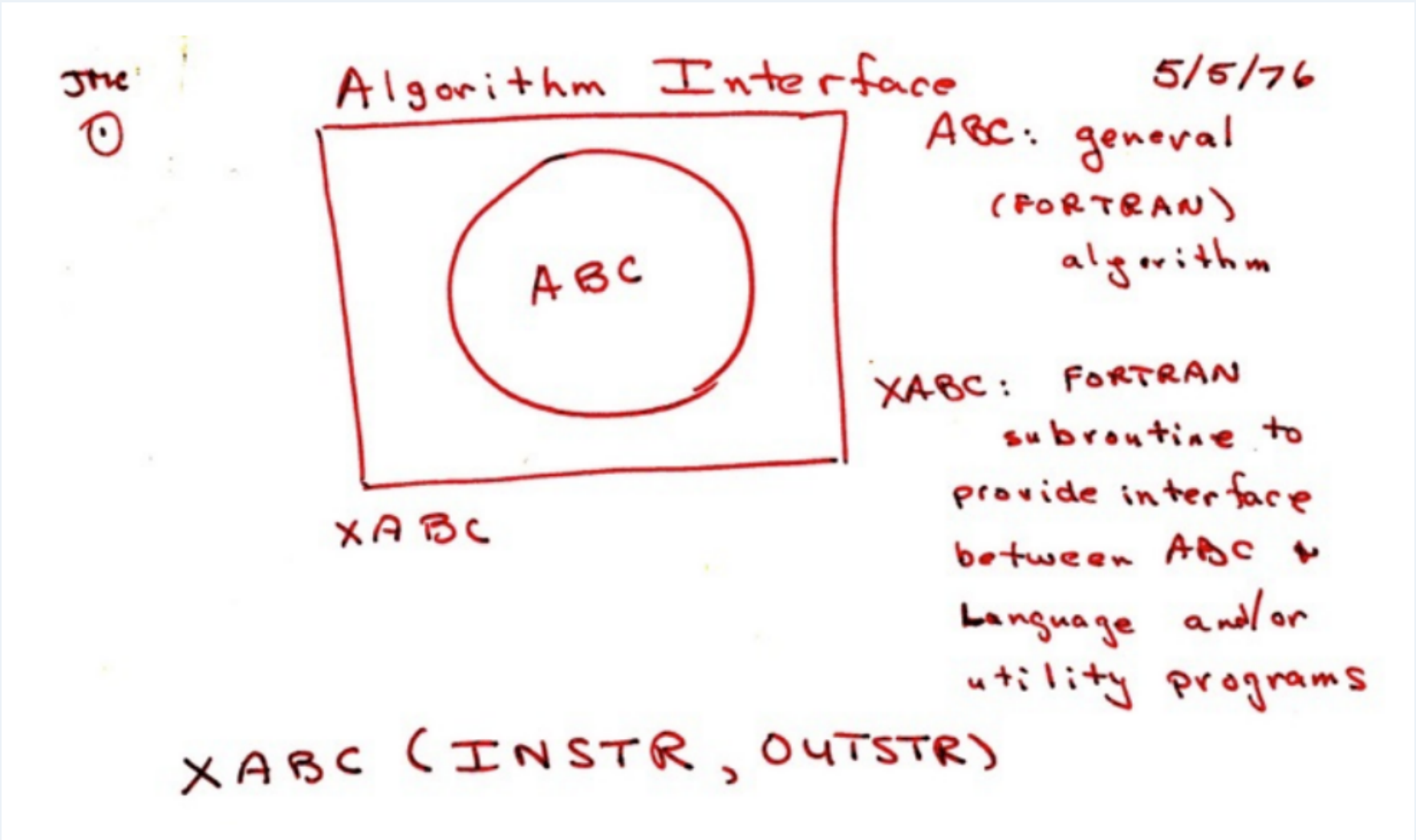
Programming with Big Data in R (pbdR)

- Free/Open Source R packages.
- Actively maintained, available on CRAN and Github.
- Packages: Library interfaces, high-level frameworks, profiling and vis, ...
- For today: MPI+ScaLAPACK stuff.
- Syntax *identical* to R.

About the Name

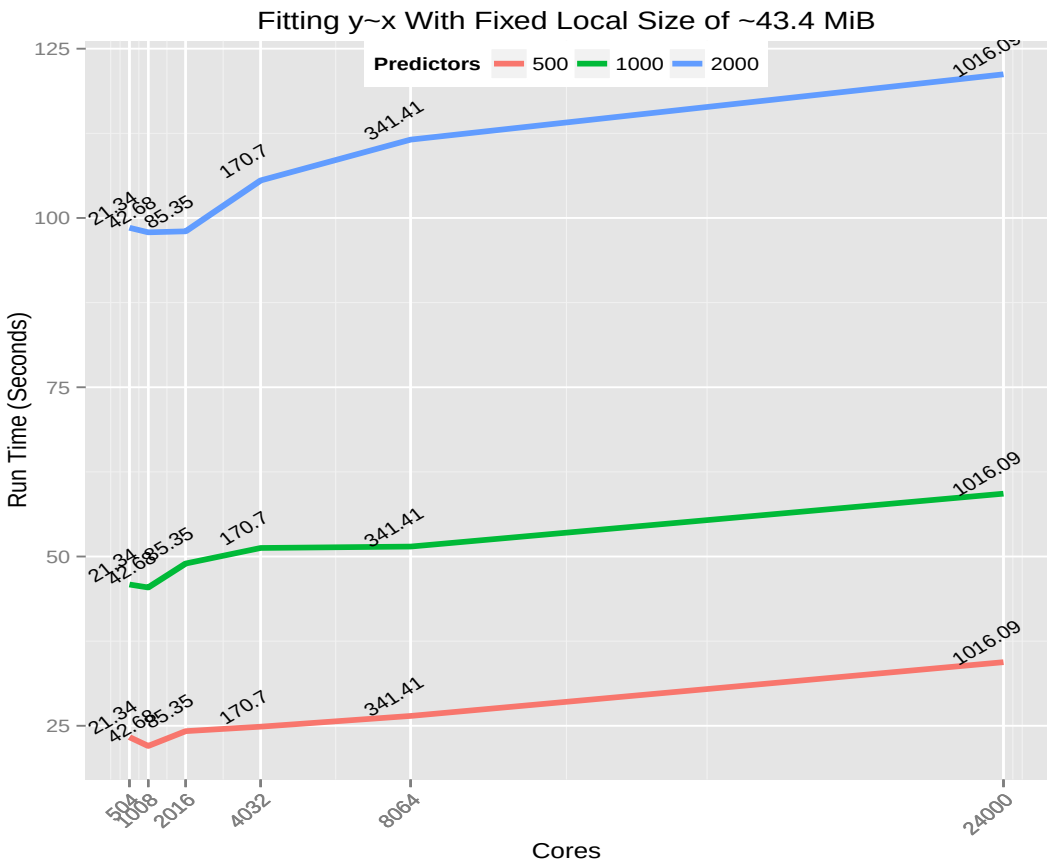


R as an Interface Language



http://datascience.la/john-chambers-user-2014-keynote/

Distributed Matrices and Statistics with **pbdDMAT**




```
library(pbdDMAT)
init.grid()

x <- ddmatrix("rnorm",
    nrow=m, ncol=n)
y <- ddmatrix("rnorm",
    nrow=m, ncol=1)

mdl <- lm.fit(x=x, y=y)

finalize()
```

pbdR Paradigms/"Opinions"

- In core.
- Focus on dense data.
- Integrates well with traditional HPC.
- C and Fortran for performance; R for the interface.
- Exclusively batch.
- (Mostly) Very different from Hadoop/Spark!

"OLCF Researchers Scale R to Tackle Big Science Data Sets"

HPC wire

- A problem that takes several hours on Apache Spark [was analyzed] in less than a minute using R on OLCF high-performance hardware.
- "...for situations where one needs interactive near-real-time analysis, **the pbdR approach is much better**."

https://www.hpcwire.com/2016/07/06/olcf-researchers-scale-r-tackle-big-science-data-sets/

1 Background and Motivation

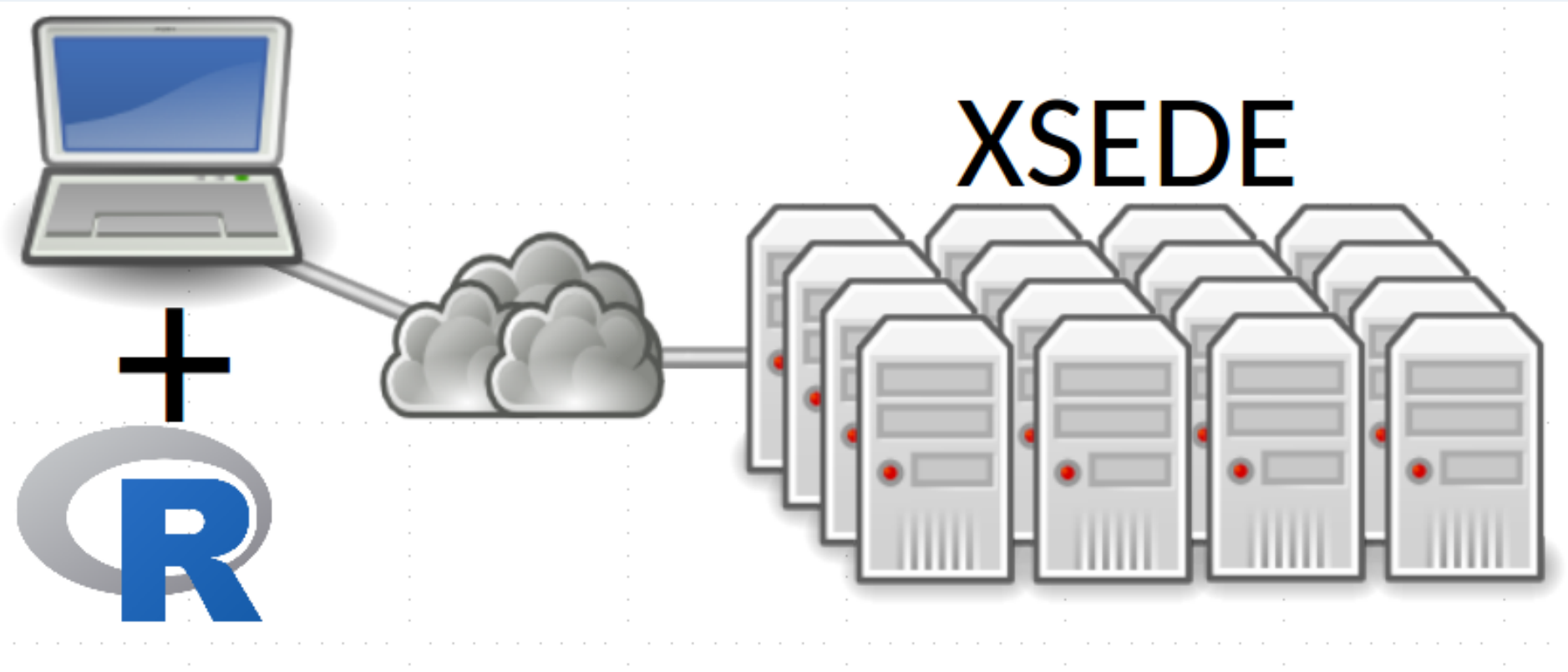
2 pbdR Compute Backend

3 The Client/Server
- Design and Architecture of the Client Server
- Examples
- Overhead

4 Challenges and Future Work

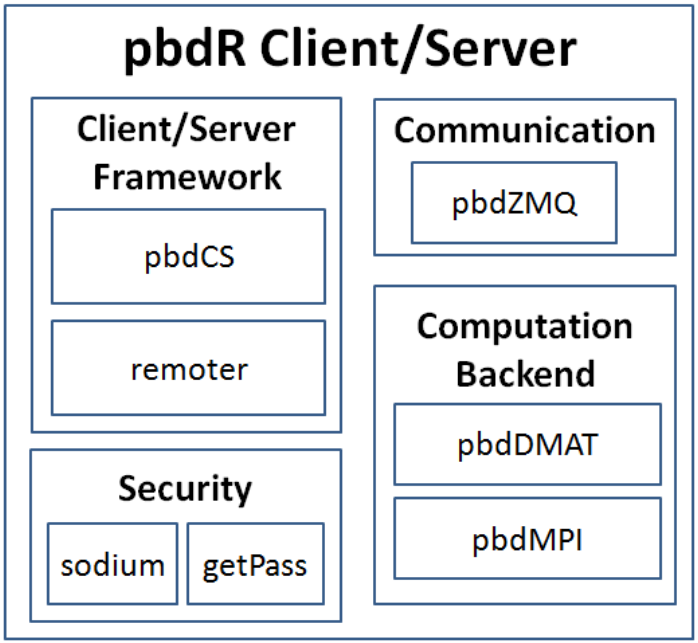
3 The Client/Server
- Design and Architecture of the Client Server
- Examples
- Overhead

Motivation



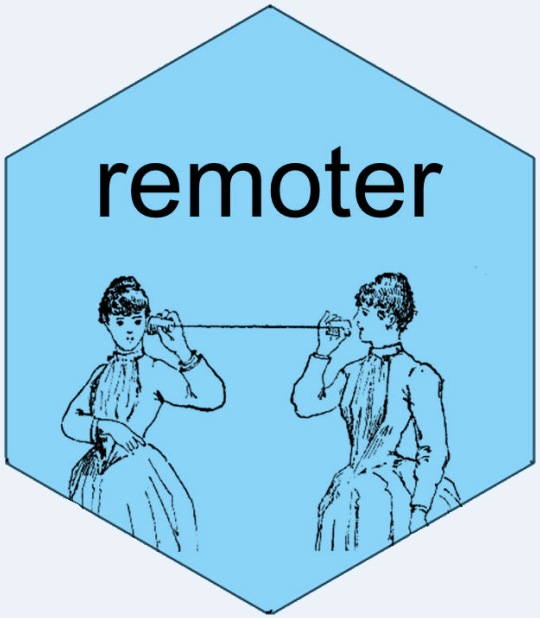
Connect local R session to a remote one.

The C/S Framework



- No specialized software environment; just R packages.
- Designed for the cloud or HPC resources.
- Mostly "un-opinionated"; can use our compute backend, SparkR, whatever.
- Optional security features: passwords and encryption.

A Quick Summary of the Main Components



- **pbdZMQ**: ZeroMQ bindings.
  Use: our C/S packages; IRkernel, ...
- **remoter**: client/server core.
  Use: cloud computing, "reference" big data framework.
- **pbdCS**: Extension of remoter for the MPI backend stuff.
  Use: Bringing interactivity to pbdR "big data" system.

Comparison to Similar Utilities

"Moving computations to a remote machine"

- Web frameworks: **shiny**, **htmlwidgets**, **fiery**, ...
- ~~Revolution~~ Microsoft Analytics: Microsoft R Server and R Client
- Jupyter/notebooks.
- RStudio Server
- **rmote**

3 The Client/Server
- Design and Architecture of the Client Server
- Examples
- Overhead

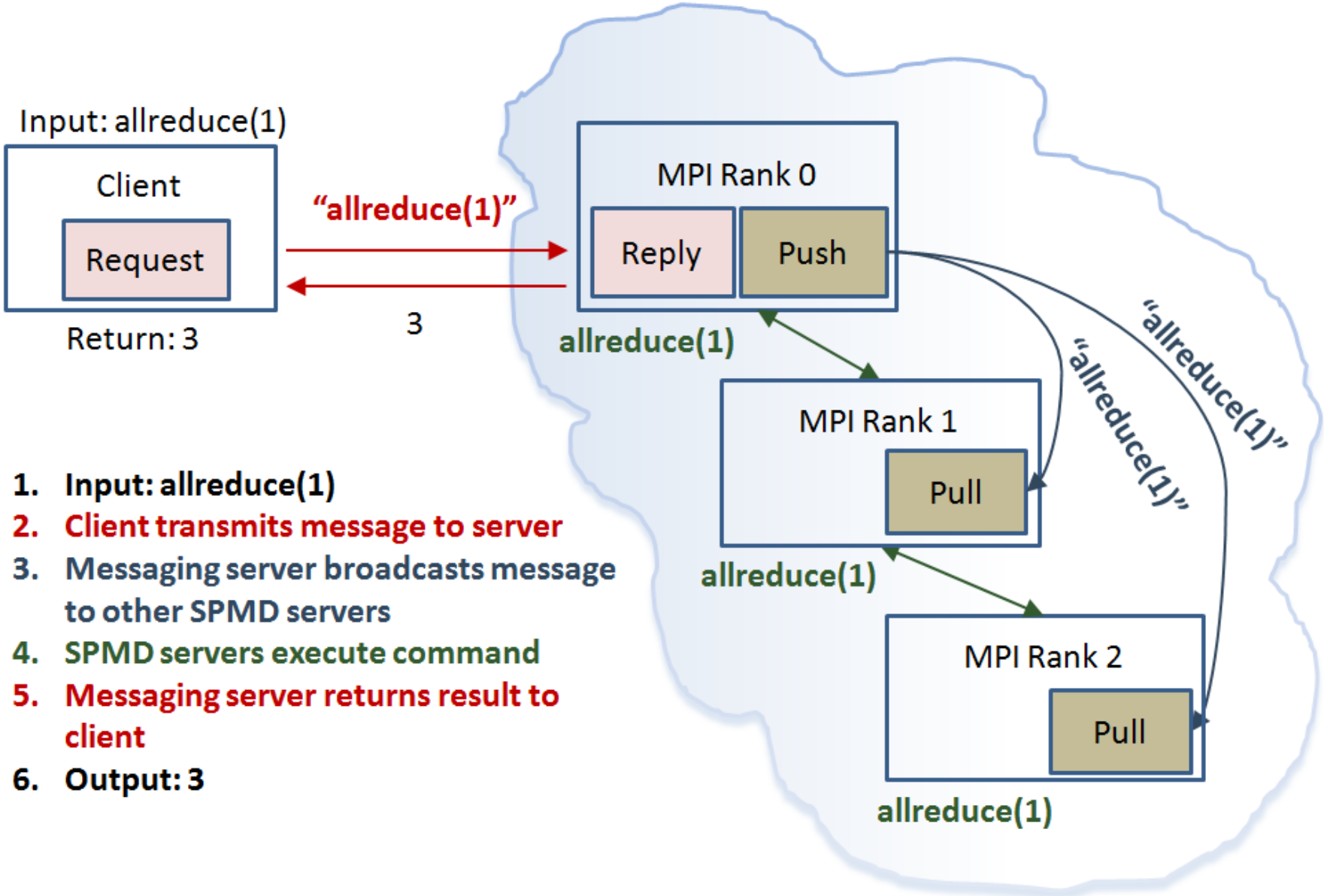
Basics

Elsewhere:

```
R> remoter::server()

## [2016-07-19 12:53:23]: *** Launching secure server ***
```

In your local R session:

```
R> remoter::client()
remoter> x = 1
remoter> x
## [1] 1
remoter> exit()

R> x
## Error: object 'x' not found
R> remoter::client()

remoter> s2c(x)
remoter> exit()
R> x
 [1] 1
```



1. **Input: allreduce(1)**
2. **Client transmits message to server**
3. **Messaging server broadcasts message to other SPMD servers**
4. **SPMD servers execute command**
5. **Messaging server returns result to client**
6. **Output: 3**

A Demo?

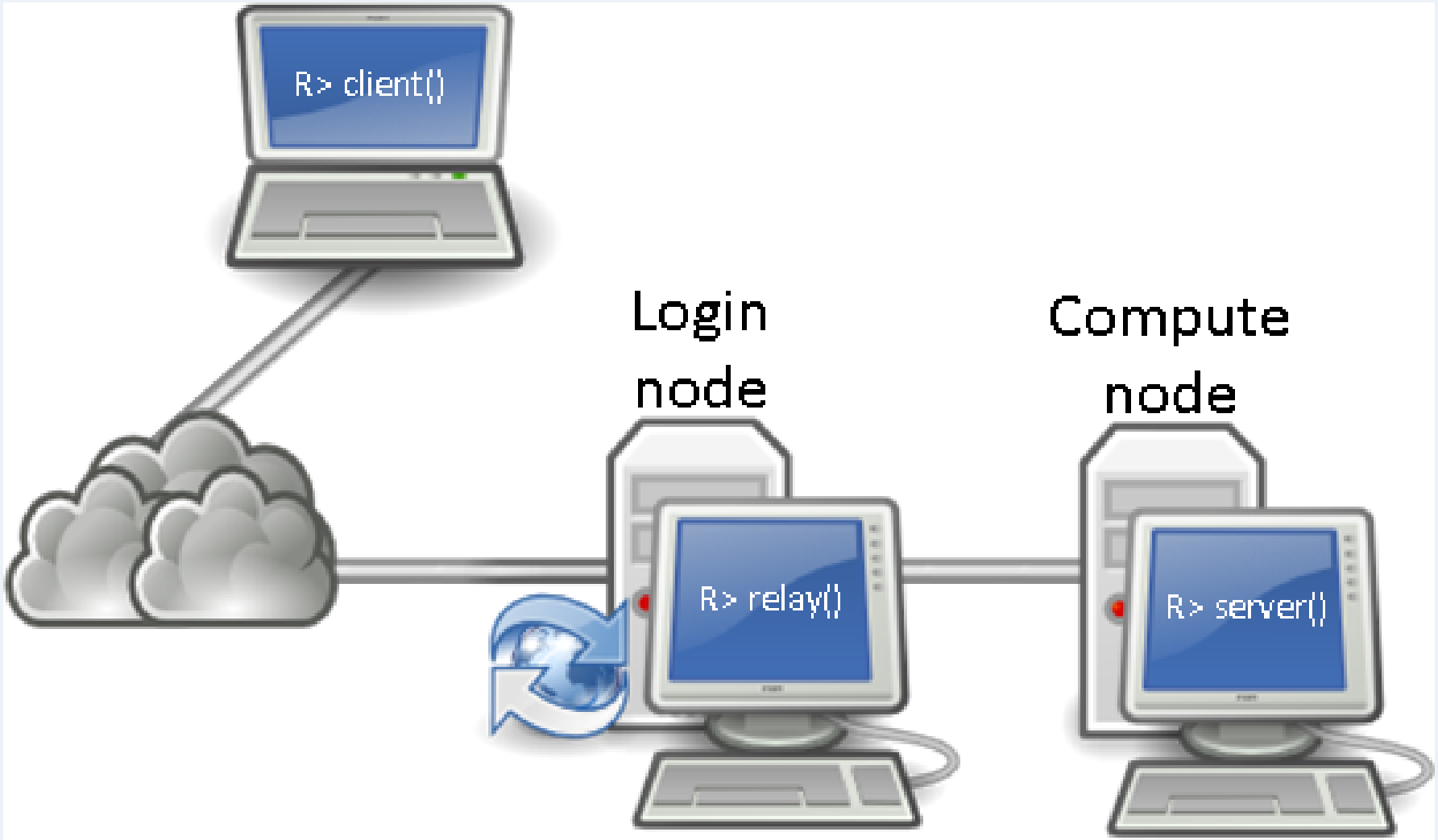
- No time for a live demo...
- Any time you see me at XSEDE16, ask and I'll give you one!
- For a non-live demo, see:
  https://github.com/snoweye/user2016.demo
- More information in the **remoter** vignette:
  https://cran.r-project.org/web/packages/remoter/vignettes/remoter.html

3 The Client/Server
- Design and Architecture of the Client Server
- Examples
- Overhead

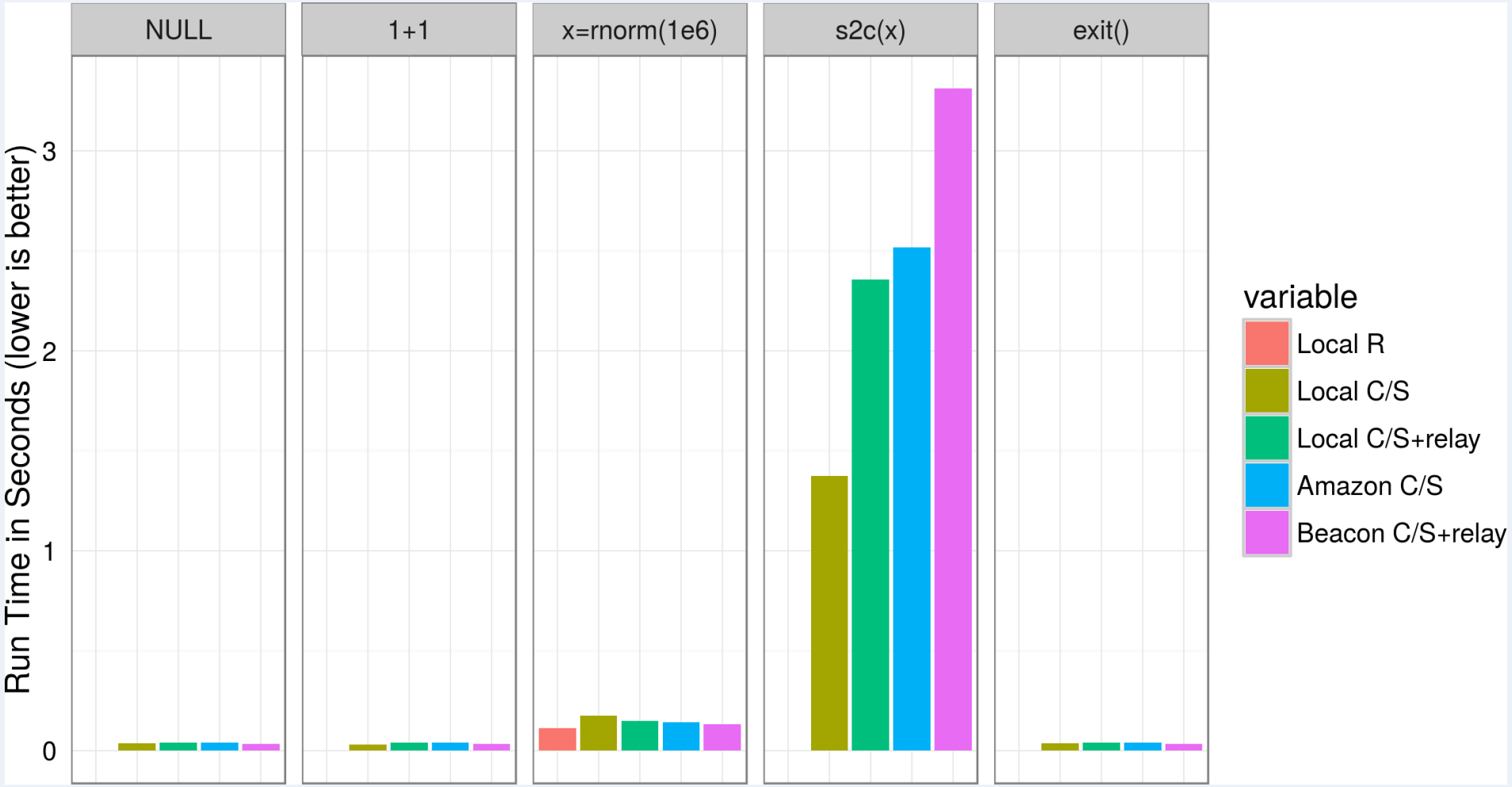
Size Costs

- There is a size overhead in transmitted messages.
- For almost everything you would do, no big deal.
- See paper for full discussion.

Relays



Run Time Costs



1 Background and Motivation

2 pbdR Compute Backend

3 The Client/Server

4 Challenges and Future Work

Already done!

- Batch computing with a remote server.
- Graphics and help, integration with **rmote**.
- Many small things we didn't bother to mention.

Object Guarding

Currently possible to send huge amounts of data without realizing it.

```
pbdR> print(huge_thing)
## uh-oh!
```

Relays and Integration with the Scheduler

Some day...

```
R> bigJob(machine="stampede", nodes=10, interactive=TRUE)
```

References

- Schmidt, D., Chen, W.-C. and Ostrouchov, G. (2016), "Introducing a New Client/Server Framework for Big Data Analytics with the R Language, XSEDE2016.
- Chen, W.-C., Ostrouchov, G., Schmidt, D., Patel, P. and Yu, H. (2012), "A Quick Guide for the pbdMPI Package. R Vignette, URL http://cran.r-project.org/package=pbdMPI
- Schmidt, D., Chen, W.-C., Patel, P., and Ostrouchov, G. (2014), "Speaking Serial R with a Parallel Accent. R Vignette, URL http://cran.r-project.org/package=pbdDEMO
- Chen, W.-C., Ostrouchov, G., Pugmire, D., Prahat, and Wehner, M. (2013), "A Parallel EM Algorithm for Model-Based Clustering Applied to the Exploration of Large Spatio-Temporal Data, Technometrics, 55(4), 513-523.

~Thanks!~

Questions?


Email: wrathematics@gmail.com


GitHub: https://github.com/wrathematics


Web: http://wrathematics.info


Blog: http://librestats.com


Twitter: @wrathematics