# R与高性能运算

## 以及R在中国未来发展的讨论

李舰

Email: lijian.pku@gmail.com

Homepage: www.leejian.name

第三届中国R语言会议（上海分会）

2010 年 11 月

# 目 录

# 目 录

# 目 录

# 目 录

# 目 录

- S语言的一个设计理念
  - Computer time is inexpensive in comparison with personnel time
  - 人的时间永远比机器的时间宝贵
- R的一个经常被质疑的弱点
  - R itself does not allow parallel execution. There are some existing solutions...

# 影响性能的因素及解决办法

- 内存不足
  - 升级内存，换用64位系统
  - 合理的内存管理
  - 大规模数据的处理
- 复杂的任务
  - 优化算法，改变思路
  - 升级CPU，使用GPU
  - 并行计算
- 计算能力不足
  - 调用C或者Fortran
  - 优化代数运算

# 内存限制

- 32位系统
  - 操作系统能支持的最大内存为$\frac{2^{32}}{1024^3} = 4G$
  - R在32位Windows系统中最大使用内存为$3G$
  - 通常R在内存使用超过$2G$（Windows）或$3G$（Unix）时就会报错
- 64位系统
  - 操作系统能支持的最大内存为$\frac{2^{64}}{1024^3} = 172 \times 10^8 G$
  - 32位的R在一些64位的Windows系统下最大内存也只有$4G$
  - R能使用的内存受限于硬件和操作系统
  - 由于没有64位的整型数据结构，无法定义过大的矩阵

# cannot allocate vector of size

- memory.size(max = FALSE)
  - memory.size(NA)，查看系统分配给R的最大内存
  - memory.size(F)，查看当前已经使用的内存
  - memory.size(T)，查看已分配的内存
- memory.limit(size = NA)
  - memory.limit()，查看系统分配给R的最大内存
  - memory.limit(2000)，分配最大内存为2G
- Rgui –max-mem-size 2Gb

# 了解R的内存

- 内存划分
  - 堆内存（Heap），基本单元是“Vcells”，每个大小为8字节
  - 地址对（cons cells），最小单元一般在32位系统中是28字节、64位系统中是56字节
- R的存储模式
  - storage.mode（x）
  - 对于整型矩阵，storage.mode(x）$<-$ "integer"
- R对象内存
  - ls()，查看当前对象
  - object.size(x)，查看对象所占用的内存

# R的垃圾清理

- rm()和gc()
  - rm()清除对象的引用
  - gc()清除内存空间，进行垃圾清理
- 程序中的垃圾清理
  - 对于长度增加的矩阵，尽量先定义一个大矩阵，然后逐步增加
  - 时刻注意清除中间对象

# BLAS和LAPACK

- BLAS
  - Basic Linear Algebra Subprograms，基础线性代数程序集
  - http://www.netlib.org/blas/
  - 基础的线性代数操作，向量和矩阵乘法等
  - C. L. Lawson, R. J. Hanson, D. Kincaid, and F. T. Krogh, Basic Linear Algebra Subprograms for FORTRAN usage, ACM Trans. Math. Soft., 5 (1979)
- LAPACK
  - Linear Algebra PACKage
  - 依赖BLAS，解多元线性方程式、线性系统方程组的最小平方解、计算特征向量、用于计算矩阵QR分解的Householder转换、以及奇异值分解等

# 优化BLAS

- BLAS的实现
  - NetLib维护的开源版本
  - Intel，MKL
  - AMD，ACML
  - NVIDIA，CUDA
  - Kazushige Goto，GOTOBLAS
- ATLAS和R的支持
  - netlib上优化BLAS的工程，如今已移到http://math-atlas.sourceforge.net/
  - 针对不同硬件环境优化
  - http://mirrors.geoexpat.com/cran/bin/windows/contrib/ATLAS/
- 操作方式
  - 下载Rblas.dll
  - 替换%R_HOME%下bin中的Rblas.dll

# bigmemory相关package

- bigmemory
  - 建立基于文件的大矩阵，实现nwitch、order等方法
- biganalytics
  - 对于bigmemory对象，实现apply、kmeans、lm、colmean等方法
- bigtabulate
  - 实现大矩阵的table、tapply等方法
- synchronicity
  - 实现Boost mutex的功能
- bigalgebra
  - 实现大矩阵的BLAS和LAPACK
  - 正处于万众期待中

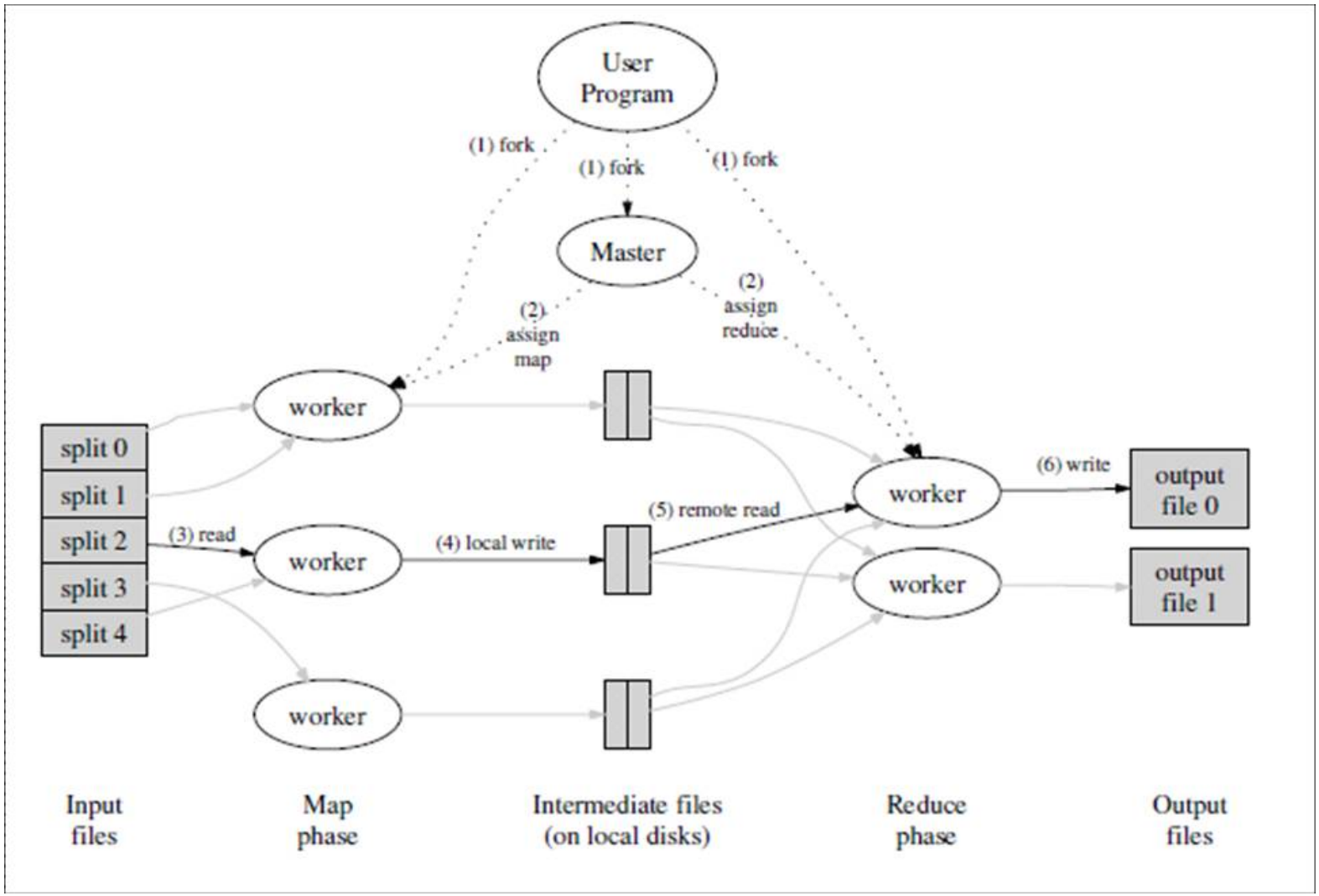
# 适用范围

- 和ff包的比较
  - 功能差别不大
  - 目前来看运行效率差不多
  - 暂无代数运算函数库的计划
- 效率实测
  - 对10G以下数据支持得很好
  - 10G以上的数据最好使用MapReduce

# MapReduce

- Google的一个专利申请
  - 2010年1月获批，编号为7 650 331，名为System and method for efficient large-scale data processing（高效大规模数据处理）。是Google最引为自豪的成果之一，也是云计算最重要的核心技术之一。
- MapReduce的应用
  - Google基础应用
  - 雅虎搜索
  - Amazon的Elastic MapReduce服务
  - 开源项目Apache Hadoop

# Google的MapReduce执行方式



# R Package: mapReduce

- MapReduce思路的简单实现
  - 基于apply系列函数
  - 功能和by以及aggregate相似
- 实现方式
  - 使用split函数将矩阵进行拆分
  - 使用apply函数并行处理
  - 汇总输出

# RHIPE简介

- 开源的MapReduce：Hadoop
  - Hadoop 是Google MapReduce 的一个Java实现
  - 定义Mapper，处理输入的Key-Value对，输出中间结果。定义Reducer，可选，对中间结果进行规约，输出最终结果。定义main函数。
  - 提交JOB，系统自动完成
- R和Hadoop的整合：RHIPE
  - 开源项目，将R和Hadoop集成在一起
  - 目前只有Linux和Mac OS版本

# R与数据库

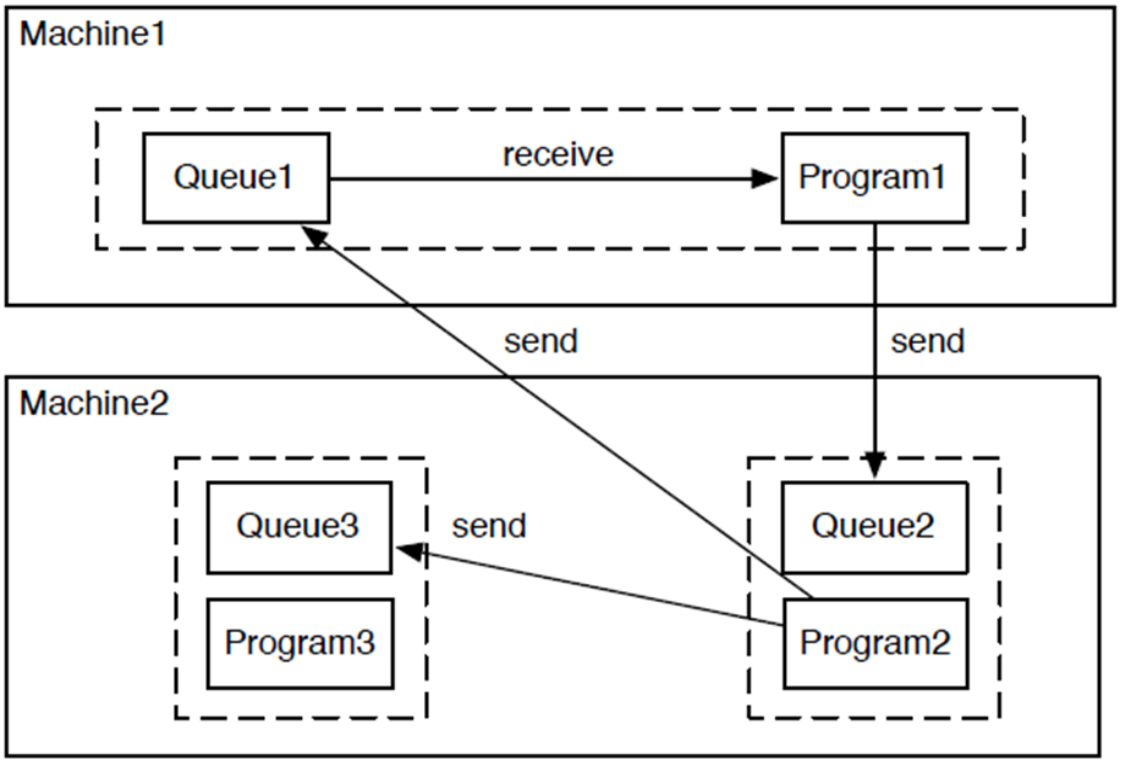
- DBI
  - RSQLite
- rodbc

# 基础知识

- 并行计算
  - 同时进行不同的计算
  - 单核多线程
  - 多核或多CPU的并行
- 显式并行
  - 由用户控制的并行，需要在算法上做专门的处理
- 隐式并行
  - 系统自动进行的并行处理

# MPI

- 什么是MPI
  - Message Passing Interface，一种消息传递接口
  - 程序通过收发队列里的消息进行交互
- MPI实现机制



# snow家族

- snow
  - 可以使用MPI、NWS、PVM、Sockets四种传递方式进行并行
  - 在多核或者计算机集群上实现并行计算
- snowfall
  - 对snow进行简化包装后的一个包
- http://www.wentrue.net/blog/?p=878

# 隐式并行

- multicore
  - 通过类似lapply的方式拆分任务
  - 自动分配到多个核计算
- doMC和foreach
  - doMC注册多核
  - foreach取代循环，分配到多核运算

# 关于R在中国未来发展的讨论

- R语言在国内传播的印象
- chinaR会议回顾
- R逐渐深入各个应用领域
- 一些倡议
  - 稳定的交流平台
  - 定期的交流讨论

# Thank you!

Email: lijian.pku@gmail.com
Homepage: www.leejian.name