Paralelização da Linguagem R em Ambientes Multicore Aplicada à Modelos de Simulação de Culturas

Paloma Rizzi, Everton de Matos,
Andriele Busatto do Carmo, Carlos Amaral Hölbig

Curso de Ciência da Computação – ICEG
Universidade de Passo Fundo (UPF)
99.001-970 – Passo Fundo – RS – Brasil
{101082,111950,andriele,holbig}@upf.br

Resumo. *Modelos de simulação do crescimento e desenvolvimento de culturas têm sido usados com sucesso ao redor do mundo na agricultura para aumentar a produtividade e reduzir custos. Uma das tecnologias que estão sendo utilizadas atualmente para implementar estes modelos é a linguagem R. Entretanto, com a crescente complexidade dos modelos, torna-se cada vez mais necessários rodar estes modelos em ambientes computacionais paralelos. Devido a isto, este artigo visa o uso da linguagem R em ambientes computacionais multicore visando a paralelização de modelos de simulação de culturas de plantas e doenças.*

1. Introdução

Modelos de simulação do crescimento e desenvolvimento de culturas têm sido usados com sucesso ao redor do mundo na agricultura para aumentar a produtividade e reduzir custos [PAVAN 2007]. Uma das tecnologias que estão sendo utilizadas atualmente para implementar estes modelos é a linguagem R [Adler, 2009; R-Project, 2010]. Trata-se se uma suíte de softwares integrados que proporcionam facilidades na manipulação de dados, no uso de funções estatísticas e na geração de gráficos. Entretanto, com a crescente complexidade dos modelos, torna-se cada vez mais necessário rodar estes modelos em ambientes computacionais paralelos. Devido a isto, este trabalho foca o uso da linguagem R em ambientes computacionais multicore [Schmidberger, et al., 2009] visando a paralelização de modelos de culturas. Esta paralelização está sendo realizada por meio do uso do pacote Multicore [Multicore, 2011], um pacote para a linguagem R que possibilita uma maneira de executar em paralelo programas implementados em R para máquinas multicore.

2. Metodologia

A linguagem R é um projeto *open source* que está disponível para a maioria das plataformas computacionais. Além de ser uma linguagem de programação também é um ambiente para computação estatística, modelagem e visualização de dados [Adler, 2007]. Por estas características pretende-se integrá-lo com o núcleo dos modelos de cultura que estão implementados na linguagem Fortran. Estes modelos apresentam um

constante aumento de dados tornando os problemas muito complexos e demandando um grande esforço computacional, o que eleva muito o tempo de processamento dos modelos. Para ser viável trabalhar com este grande número de dados é cada vez mais importante o uso de computação paralela. A linguagem R apresenta diversos pacotes que possibilitam a paralelização. Existem pacotes para Clusters, Grids e para máquinas multicore, sendo este último o contexto escolhido, em um primeiro momento, para a pesquisa. Alguns destes pacotes serão instalados, analisados e comparados por esta pesquisa. Uma lista atual destes pacotes poderá ser encontrada em http://cran.r-project.org/web/views/HighPerformanceComputing.html (CRAN Task View: High-Performance and Parallel Computing with R) que é a página oficial da entidade que disponibiliza o R e seus pacotes oficiais.

3. Resultados e Discussões

Existem diversos pacotes para programação paralela em R, por esse motivo foi feito um estudo para comparar os pacotes e decidir qual a melhor opção. Para clusters os principais pacotes utilizados são Rmpi, e Rpvm, respectivamente invólucros de MPI e PVM para uso em R. O RPVM, R Parallel Virtual Machine, é projetado para permitir uma rede de Unix heterogêneas ou máquinas Windows a serem usadas como um único computador paralelo distribuído. O RPVM é complexo de ser utilizado por valer-se de funções de baixo nível. Em cluster PVM foi largamente utilizado, entretando vem perdendo espaço para o MPI que está se tornando padrão na computação paralela. RMPI, Message-Passing Interface, é um sistema padronizado e portátil de transmissão de mensagens em computação paralela, fornecendo uma interface R para funções MPI de baixo nível. Desta forma, o utilizador R não precisa conhecer os detalhes das implementações de MPI. Assim, para clusters o RMPI apresenta melhor solução que o RPVM. Alguns destes pacotes são apresentados na Tabela 1.

Tabela 1 - Overview sobre computação paralela com R em clusters de computadores

Pacote	Descrição
rpvm	Interface R para PVM (Parallel Virtual Machine)
Rmpi	Interface (Wrapper) para MPI (Message-Passing Interface)
snow	**S**imple **N**etwork **o**f **W**orkstations
snowFT	Pacote de tolerância a falhas para o snow
snowfall	Fácil computação em cluster baseada no snow
papply	Função apply em paralelo usando MPI
taskPR	Pacote R task-parallel
foreach	Construtor foreach para R
doMC	Foreach paralelo adapatado para o pacote multicore
doSNOW	Foreach paralelo adapatado para o pacote snow
doMPI	Foreach paralelo adapatado para o pacote Rmpi
Rdsm	Ambiente tipo threads para R

Fonte: Adaptado de [Schmidberger, 2011].

Para ambientes computacionais com processadores multicore os principais pacotes são Fork e Multicore. O pacote Fork utiliza basicamente os recursos do sistema UNIX para efetuar a paralelização. Possui uma utilização relativamente simples por ter poucas funções mas não apresenta suporte para funções de alto nível como aplly(). O pacote multicore apresenta além das chamadas de funções do sistema UNIX, outras rotinas próprias. Sua utilização é mais complexa por apresentar mais funções, porém tem suporte para funções de alto nível. Para ambientes multicore, o pacote Multicore fornece melhores soluções pois reune mais recursos que o pacote Fork. Alguns destes pacotes são apresentados na Tabela 2.

A utilização de cluster possui um custo maior que a utilização de computadores com processador multicore, além de terem uma implementação mais complexa e uma usabilidade menor. Com a popularização dos multicores, optou-se por este tipo de paralelização dos modelos de cultura, utilizando o pacote Multicore como meio para isto.

Tabela 2 - Overview sobre computação paralela com R em ambientes multicore

Pacote	**Descrição**
fork	Funções em R para manipulação de múltiplos processos
multicore	Código para processamento paralelo do R em máquinas com múltiplos cores ou CPUs

Fonte: Adaptado de [Schmidberger, 2011].

Além destes pacotes para clusters e multicore há pacotes para Grid de computadores (Tabela 3), GPUs e pacotes para aplicações específicas que podem ser obtidas no site da CRAN.

Tabela 3 - Overview sobre computação paralela com R em Grids de computadores

Pacote	**Descrição**
gridR	Executa funções em hosts remotos, clusters ou grids

Fonte: Adaptado de [Schmidberger, 2011].

4 Conclusões e Trabalhos Futuros

Com os modelos de simulação apresentando cada vez mais dados é imprescindível encontrar formas de otimizar o desempenho dos mesmos. Para tanto, a paralelização mostra ser uma das alternativas, pois pode melhorar efetivamente o tempo de execução dos programas. O pacote Multicore vem ao encontro das necessidades dos modelos de simulação viabilizando a programação concorrente destes.

Referências Bibliográficas

Adler, J. (2009) R in a Nutshell. 1. ed. Sebastopol: O'Reilly.

Multicore. (2011) Package MULTICOR. Disponível em: <http://www.rforge.net/multicore/>. Acesso em 5 ago. 2011.

Pavan, W. (2007) Técnicas de engenharia de software aplicadas à modelagem e simulação de doenças de plantas, 2007. 182 p. Tese (Doutorado em Agronomia)- Faculdade de Engenharia e Medicina Veterinária, Universidade de Passo Fundo, Passo Fundo.

R-PROJECT. (2010) **R Project for Statistical Computing**. Disponível em: <http://www.r-project.org/>. Acesso em 6 Dez. 2010.

Schmidberger, M.; Morgan, M.; Eddelbuettel, D.; Yu, H.; Tierney, L.; Mansmann, U. (2009) State of the Art in Parallel Computing with R. Journal of Statistical Software, vol. 31, Issue 1, p. 1-27, 2009.