Boinc + R : Executando rotinas de bioinformática em grades oportunistas

Rodrigo L. M. Flores
Orientador: Roberto Hirata Jr.
Instituto de Matemática e Estatística
Universidade de São Paulo
{flores, hirata}@ime.usp.br

Palavras chave: BOINC, R, Bioinformática, Computação em grade

Resumo

O R é um ambiente poderoso para análise e visualização de dados em bioinformática. Porém, muitas rotinas de análise são combinatórias, o que demanda muitos recursos computacionais. O objetivo deste trabalho é, utilizando uma rede de computadores do IME-USP, criar uma grade computacional para processamento destas rotinas, utilizando o renomado middleware BOINC como middleware para a distribuição dos workunit.

1. Introdução e Objetivos

Algoritmos da área de bioinformática normalmente são combinatórios e bastante custosos computacionalmente e muitas vezes seu processamento em um único computador pessoal se torna inviável e demorado. A utilização de um *supercomputador* pode ser uma boa opção, mas tais tipos de computadores são específicos e caros, o que dificulta bastante seu uso. Uma outra opção é utilizar a computação oportunista em grades de computadores de uma universidade ou empresa, já que estas normalmente possuem muitos computadores que normalmente ficariam desligados em períodos fora do expediente ou, na maior parte do tempo, sub-utilizados no período de expediente.
Como já existem muitas bibliotecas para desenvolvimento de aplicações de bioinformática na linguagem *R*, um dos requisitos do trabalho é fazer com que rotinas nesta linguagem possam ser executadas na grade, evitando assim a reimplementação dos algoritmos e programas.
Para o gerenciamento da grade, utilizamos o já renomado *middleware BOINC*, utilizado em diversos projetos de computação em grade voluntária (pessoas voluntariamente doando ciclos de CPU para projetos) e que vem também sendo utilizado em diversos projetos de computação em grade como o citado em [GGdVS08].
O objetivo deste trabalho é criar uma grade de computadores na rede CEC do IME-USP utilizando o *BOINC* para distribuir as tarefas entre os computadores. A rede CEC disponibilizou cerca de 50 máquinas, tanto com sistemas Linux ou Windows para o cliente ser instalado e a rede funcionar como uma grade de computadores.

2. Metodologia

O trabalho foi baseado no artigo [GGdVS08], que contava da experiência do uso do *BOINC* em uma universidade da Espanha para processamento em grade. Baseado neste trabalho decidimos utilizar o *BOINC* para o gerenciamento de nossa grade. Um outros trabalhos que serviu de inspiração foram o [RAD09], cujo objetivo foi instalar uma rede parecida com a nossa baseada no *middleware Alchemi* e na plataforma *.NET* e o artigo [SME+09] que fez um visão geral das alternativas de computação de alta performance para a linguagem *R* também nos fez ter uma visão geral sobre as alternativas de computação de alta performance com o *R*.
Como a API do *BOINC* funciona para poucas linguagens, foi necessária a utilização de um programa *wrapper* que executa um arquivo binário seguindo as instruções fornecidas em um arquivo XML. Para executar as rotinas em R, colocamos como parâmetro um outro programa que apenas chama a função do C, system e usando como parâmetro o caminho do interpretador junto com o script.

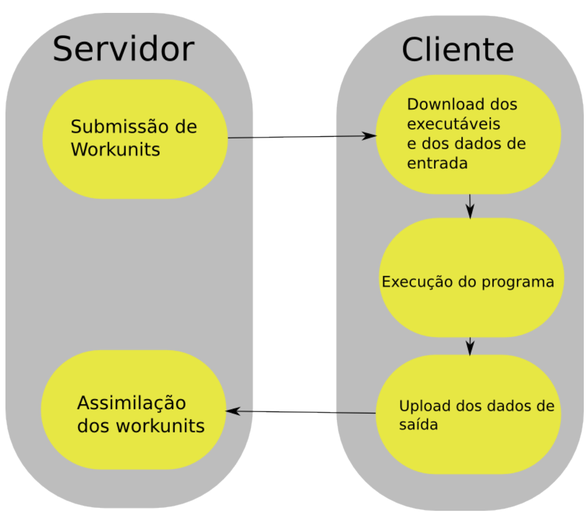

Figura 1: *Funcionamento do BOINC*

3. Funcionamento do BOINC com a linguagem *R*

O *BOINC* funciona da seguinte maneira: criado um *workunit* que é um conjunto de dados de entrada para uma aplicação, o cliente se conecta no servidor, faz o download dos arquivos executáveis da aplicação e dos dados de entrada. Após isso, os arquivos são executados com a entrada. Quando o processamento acaba, o cliente faz o upload da saída e o servidor faz a assimilação da saída, colocando as informações em um banco de dados. O funcionamento do Boinc pode ser visto na figura 1. O *wrapper* funciona na etapa de processamento: seguindo uma configuração em um arquivo XML, o *wrapper* executa um programa que executa o interpretador R junto com o script em *R*.

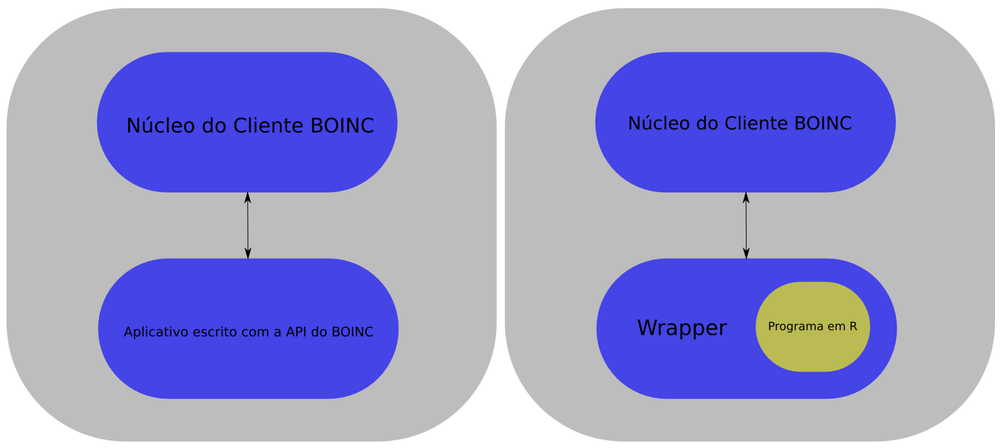

Figura 2: *Funcionamento do BOINC com um aplicativo usando a API e com o Wrapper*

4. Resultados e discussão

A parte do servidor já está implementada e em funcionamento. A instalação dos clientes está em fase de implantação na Rede CEC do IME-USP e em breve a grade estará em seu pleno funcionamento.
A instalação do servidor foi simples: havia um bom guia de instalação e configuração para o *Debian GNU/Linux*. Porém, a maior dificuldade foi conseguir executar os programas em *R* já que haviam bugs na configuração de compilação do *wrapper* para o Windows, e o *BOINC* não possuía mensagens de erro que nos permitiam facilmente encontrar o problema.

5. Conclusão

O objetivo está bem próximo de ser concluído: só é necessário terminar a implantação na rede e fazer o anúncio pedindo trabalhos para serem executados na grade. Outro ponto interessante é que é possível ajudar no processamento usando tanto máquinas com sistemas Linux quanto Windows o que não foi feito nos outros trabalhos semelhantes citados no projeto.
Esperamos que a grade seja de grande ajuda a grupos de bioinformática que precisem de recursos computacionais para fazer análises de dados.

Referências

[GGdVS08] D.L. Gonzalez, G.G. Gil, F.F. de Vega, and B. Segal, *Centralized boinc resources manager for institutional networks*, April 2008, pp. 1–8.

[RAD09] Roberto Hirata Jr. Rodrigo A. Dias, *Middle-r - a user level middleware for statistical computing*, VII Workshop on Grid Computing and Applications (2009).

[SME+09] Markus Schmidberger, Martin Morgan, Dirk Eddelbuettel, Hao Yu, Luke Tierney, and Ulrich Mansmann, *State-of-the-art in parallel computing with r*, 1 2009.