

Proposta de Trabalho de Conclusão de Curso

Rodrigo L. M. Flores

#USP:5127470

Orientador: Prof. Dr. Roberto Hirata Jr.

15 de junho de 2009

Sumário

1	Integrantes do trabalho	2
2	Tema da monografia	2
3	Resumo da monografia	2
4	Objetivos	2
5	Atividades realizadas	3
6	Cronograma	4
6.1	Julho	4
6.2	Agosto	4
6.3	Setembro	4
6.4	Outubro	4
6.5	Novembro	5
7	Estrutura esperada da monografia	5
7.1	Introdução	5
7.2	Conceitos e tecnologias estudadas	5
7.3	Processamentos realizados	5
7.4	Discussão	5
7.5	Conclusão	6
7.6	Parte subjetiva	6

1 Integrantes do trabalho

O trabalho será feito individualmente por Rodrigo Luiz Marques Flores, orientado pelo Prof. Dr. Roberto Hirata Jr. .

2 Tema da monografia

Os temas da monografia serão:

- Bioinformática, mais especificamente na área de reconhecimento de padrões e aprendizado computacional.
- Computação em grade aplicada à ferramenta estatística *R*.

3 Resumo da monografia

O *R*¹ é uma ferramenta estatística muito utilizada na área de bioinformática, tanto que já existem soluções integradas com essa ferramenta, como o Bioconductor² para o auxílio do pesquisador desta área. Como muitos algoritmos desta área são combinatórios, o processamento em um computador acaba sendo bastante demorado e a utilização de *clusters* ou grades de computadores se faz necessária. Devido ao alto custo de um cluster e a disponibilidade de um parque de máquinas de uso cotidiano em uma universidade ou empresa, a utilização de grades de computadores acaba sendo mais viável.

O projeto terá como objetivo a instalação de uma grade de computadores na rede *CEC* do IME-USP que deverá ser capaz de processar distribuidamente rotinas na linguagem de programação do *R*.

4 Objetivos

Os principais objetivos deste trabalho de conclusão de curso são:

- Estudar os diversos middlewares disponíveis e estabelecer a comparação entre eles, utilizando sempre como requisito o processamento de rotinas do *R*;

¹Disponível em <http://www.r-project.org>

²Disponível em <http://www.bioconductor.org>

- Estudar a performance do processamento em máquinas com sistemas operacionais *GNU/Linux* e em máquinas com sistemas operacionais *Microsoft Windows*;
- Disponibilizar como um recurso para a comunidade científica do IME e da USP, principalmente o grupo de bioinformática, a grade de computadores;
- Estabelecer uma comparação entre o custo benefício do uso de uma grade ou da contratação de um serviço de computação em nuvem como o da empresa norte americana *Amazon*.

5 Atividades realizadas

Até o momento foram realizadas as seguintes atividades:

- **Visita a uma grade computacional:** Para um contato inicial com o assunto, foi realizada uma visita à grade computacional instalada no Colégio Rainha da Paz, que se localiza no bairro da Lapa na cidade de São Paulo. O administrador desta grade é o aluno de mestrado do IME Rodrigo Assirati Dias, que além de explicar o funcionamento, deu algumas recomendações sobre o assunto;
- **Obtenção de um computador para ser o manager da grade:** Como o manager de uma grade é muito requisitado pelos outros computadores, foi necessário utilizar um computador para uso exclusivo da grade. Este computador em nenhum momento ficará disponível para os usuários. Foi então pedido emprestado um computador para a rede Vision do IME-USP. Esta máquina já foi formatada e já está instalada com o sistema operacional *Debian GNU/Linux*;
- **Estudo teórico de casos de uso semelhantes ao nosso:** Como existem algumas alternativas de *middleware* disponíveis, foi feito um estudo sobre algumas alternativas. O artigo [GGdVS08] fala sobre a instalação do *middleware BOINC* em uma grade de computadores de uma universidade da Espanha que usou como testes rotinas na linguagem *R*. Outro artigo estudado foi o [RAD09] que usou o *middleware Alchemi*, baseado na plataforma da Microsoft *.NET*, na rede de computadores do Colégio Rainha da Paz e da faculdade SENAC;
- **Estudo de arcabouços de computação em grade com o R:** A partir do artigo [SME⁺09] que fala sobre as diversas alternativas para tal fim, foram conhecidas as alternativas, a que cada uma se propõe, e as vantagens e desvantagens do uso de cada uma.

6 Cronograma

6.1 Julho

- Preparação do laboratório para a realização do processamento do *BOINC*;
- Realização do processamento com o *BOINC*;
- Início da análise das estatísticas do processamento do *BOINC*.

6.2 Agosto

- Finalização da análise de estatísticas do processamento do *BOINC*;
- Preparação do laboratório para a realização do processamento do *GridR + Condor*;
- Realização do processamento com o *GridR + Condor*;
- Início da análise das estatísticas do processamento do *GridR + Condor*.

6.3 Setembro

- Finalização da análise de estatísticas do processamento do *GridR + Condor*;
- Preparação do laboratório para a realização do processamento com o *Alchemi*;
- Realização do processamento com o *Alchemi*;
- Preparação de resumo para os simpósios de iniciação científica do IME e da USP;
- Início da elaboração da monografia.

6.4 Outubro

- Elaboração da monografia;
- Análise das estatísticas do processamento com o *Alchemi*;
- Entrega da monografia ao orientador.

6.5 Novembro

- Apresentação do trabalho nos eventos que este possa vir a ser aprovado;
- Elaboração da parte subjetiva da monografia;
- Preparação do pôster;
- Entrega do pôster;
- Entrega da monografia ao professor responsável pelo curso, com as correções sugeridas pelo orientador.

7 Estrutura esperada da monografia

A monografia terá uma estrutura nos seguintes moldes:

7.1 Introdução

Como toda introdução, será feita uma abordagem aos temas estudados no trabalho, com a devida contextualização e possíveis utilidades práticas. Os objetivos do trabalho estarão dispostos também nesta seção.

7.2 Conceitos e tecnologias estudadas

Nesta seção será feita uma explicação sobre as ferramentas utilizadas, sobre as tecnologias utilizadas e sobre as que poderiam ter sido utilizadas e os principais problemas que as fizeram não ser utilizadas. Serão discutidos também os conceitos estudados.

7.3 Processamentos realizados

Aqui serão discutidos os processamentos realizados, com dados de *benchmark*, as estatísticas de processamento e com os relatórios. Poderão haver também alguns gráficos com os dados de cada experimento.

7.4 Discussão

A parte de discussão estabelecerá comparações dos diversos processamentos realizados, falando das vantagens e desvantagens do uso de cada ferramenta.

7.5 Conclusão

Na conclusão, falarei sobre as expectativas do trabalho e sobre se os objetivos propostos na introdução foram cumpridos ou não.

7.6 Parte subjetiva

Nesta parte será feita uma análise das matérias relevantes para o trabalho e sobre os principais obstáculos e motivações que eu tive na graduação em ciências da computação.

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, Alfredo Goldman, *Middle-r - a user level middleware for statistical computing*, VII Workshop on Grid Computing and Applications - WCGA (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.