PIBIC / PIBITI

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

XIV JORNADA DE INICIAÇÃO CIENTÍFICA E TECNOLÓGICA

Auditório B
09 de agosto de 2019

Laboratório Nacional de Computação Científica

Av. Getúlio Vargas 333. Petrópolis, RJ.

Laboratório Nacional de Computação Científica
LNCC

MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA,
INOVAÇÕES E COMUNICAÇÕES

PÁTRIA AMADA BRASIL
GOVERNO FEDERAL

Jornada de Iniciação Científica e Tecnológica do LNCC

Petrópolis, 09 de agosto de 2019.

Laboratório Nacional de Computação Científica – LNCC

Diretor
Augusto Cesar Gadelha

Coordenação de Gestão e Administração - COGEA
Anmily Paula dos Santos Martins

Coordenação de Métodos Matemáticos e Computacionais - COMAC
Frédéric Gerard Christian Valentin

Coordenação de Modelagem Computacional - COMOD
Márcio Arab Murad

Coordenação de Pós-Graduação e Aperfeiçoamento - COPGA
Artur Ziviani

Coordenação de Tecnologia da Informação e Comunicação - COTIC
Wagner Vieira Léo

Programa Institucional de Bolsas de Iniciação Científica&
Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação
Marcos Garcia Todorov

Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq

Presidente
João Luiz Filgueiras de Azevedo

Coordenadora Geral do PIBIC/PIBITI
Lucimar Batista de Almeida

Jornada de Iniciação Científica e Tecnológica do LNCC

Comissão Interna do PIBIC/PIBITI-LNCC

Marcos Garcia Todorov
Eduardo Lucio Mendes Garcia
Helio José Corrêa Barbosa
Jack Baczynski

Avaliadores Externos

Franklin Marquezino - UFRJ

José Cristiano Pereira–UCP/GE

Apresentação

O LNCC realiza este ano a XIV Edição da Jornada de Iniciação Científica e Tecnológica, que é um fórum de divulgação das pesquisas desenvolvidas no contexto dos Programas Institucionais de Bolsas de Iniciação Científica (PIBIC) e de Bolsas de Iniciação Tecnológica (PIBITI) fomentados pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). No período de agosto de 2018 a julho de 2019, o PIBIC e PIBITI congregaram alunos de várias instituições de ensino e de diversas áreas do conhecimento. Este volume apresenta os resumos dos trabalhos desenvolvidos pelos bolsistas no período. Durante a Jornada, os trabalhos são apresentados pelos bolsistas oralmente e avaliados por um comitê científico externo.

Nesta XIV Edição da Jornada, o Comitê Externo de Avaliação do PIBIC/PIBITI tem a seguinte composição:
Prof. Franklin Marquezino - UFRJ
Prof. José Cristiano Pereira–UCP/GE

Destacamos o papel relevante do PIBIC/PIBITI do LNCC no desenvolvimento das pesquisas no LNCC e, principalmente, na formação complementar dos bolsistas, promovendo o aprimoramento do conhecimento, espírito criativo, reflexão crítica e ética. Estas características têm contribuído para suas inserções no mercado de trabalho e em programas de pós-graduação, como o PPG em Modelagem Computacional do LNCC. Este é o resultado do esforço e dedicação de todos os participantes.

Agradecimentos

Agradecemos ao CNPq pelas bolsas concedidas, à Direção do LNCC pelo apoio e à Comissão Interna do PIBIC e PIBITI no LNCC.

Agradecemos a disponibilidade e contribuição dos membros do Comitê Externo de Avaliação.

O sucesso desta Jornada, e do Programa como um todo, é o resultado da dedicação e do esforço de toda a comunidade do LNCC. Expressamos em particular nosso reconhecimento ao apoio concedido pela secretaria do PPG-LNCC e, em particular, à Sra. Roberta Machado.

Marcos Garcia Todorov
Coordenador do PIBIC/PIBITI - LNCC

Índice

Indexação Métrica em Junções Espaciais 01
Bolsista: André Muniz Demori
Orientadores:Fabio Porto e Douglas Ericson de Oliveira

Segurança e Privacidade em Redes de Anonimato 07
Bolsista: Daniel Veiga da Silva Antunes
Orientador:Fábio Borges de Oliveira

Calibração de Modelos de Incrustação de CaSO4 em Trocadores de Calor.....09
Bolsista: Fernando Henrique Pereira Cardozo
Orientador:Renato Simões

Análise de Dados de Proveniência em Aplicações de BiologiaComputacional.....14
Bolsista: Guilherme da Silva Vieira
Orientadores: Kary Ann del Carmen Ocaña Gautherot, Fábio André Machado Porto e Douglas Ericson Marcelino de Oliveira

Desenvolvimento de Módulos Paralelo-Híbrido de Bioinformática para Ambientes GPU de Supercomputação.....18
Bolsista: Guilherme Freire da Silva Dornelas
Orientadores:Carla Osthoff Ferreira de Barros,Kary Ann del Carmen Ocaña Gautherot e André Elias Rodrigues Soares

Avaliação de métodos de aprendizado de máquina em aplicações deBioinformática.....21
Bolsista: Isabela Canuto Ramos
Orientadores:Kary Ann del Carmen Ocaña Gautherot,Fábio André Machado Porto e Douglas Ericson Marcelino de Oliveira

Modelagem e controle de um manipulador robótico sujeitoa falhas.....26
Bolsista: Junior do Nascimento Xavier
Orientador: Marcos Garcia Todorov

Algoritmos Bio-Inspirados e Redes Neurais Artificiais Aplicados à Segmentação de Imagens Médicas e Biológicas.....30
Bolsista: Lucas Pampolin Laheras
Orientadores:Gilson Antonio Giraldi e Paulo Sérgio Silva Rodrigues

Uso da Computação Distribuída de Alto Desempenho para Fluidodinâmica Computacional: Um Estudo de Eficiência Energética e Desempenho.....40
Bolsista: Matheus de Oliveira Pires
Orientador: Bruno Schulze

Desenvolvimento de Estratégias Autonômicas para aEficiência Energética em Ambientes HPC.....48
Bolsista: Matheus Gritz Alves de Souza
Orientador:Bruno Schulze

A interação da cevada com os nutrientes do solo e o pulgão afídeo.....56
Bolsista: Priscila Luana Lopes de Barros Weisz
Orientador:Mauricio Vieira Kritz e Lucas dos Anjos

Avaliação de Métodos de Aprendizado em Aplicações de Saúde.....63
Bolsista: Raquel de Abreu Junqueira Gritz
Orientadores:Fábio André Machado Porto e Douglas Ericson Marcelino de Oliveira

Simulação numérica e computacional do tráfego viário 72
Bolsista: René Constancio Nunes de Lima
Orientadores:Elson M. Toledoe Regina Célia P. Leal Toledo

Computação em máquinas não-confiáveis.....79
Bolsista:Ricardo Luiz Cerqueira Júnior
Orientador:Fábio Borges de Oliveira

Modelagem de sistemas térmicos.....81
Bolsista: Thiago da Rocha Canella
Orientador: Renato Simões

Inversão de DadosSísmicos.....86
Bolsista: Vinícius Theobaldo Jorge
Orientador:Marcio Rentes Borges

Verificação de propriedades físicas em Lentes Gravitacionais em Big Data.....95
Bolsista: Viviane de Mattos Matioli
Orientador:Fabio Porto

Métodos numéricos para o escoamento bifásico em meios porosos heterogêneos em ambientes computacionais de arquitetura de memória híbrida.....99
Bolsista: Weber Guilherme Dias Ribeiro
Orientadores:Carla Osthoff Barros e Frederico Luís Cabral

Bolsistas PIBITI

Inicialização de Fluidos para Animação Computacional.....105
Bolsista: Allan Carlos Amaral Ribeiro
Orientador: Gilson Antonio Giraldi

Aplicação das Ferramentas Intel Parallel Studio para modernização de código para métodos numéricos de diferenças finitas para solução de equações diferenciais parciais em arquitetura Intel Haswel/Broadwell..... 111
Bolsista: Gabriel Pinheiro da Costa
Orientador: Carla Osthoff e Frederico Luis Cabral

Técnicas de Aprendizagem de Máquina na Segmentação de Imagens Médicas de Ultrassom Intravascular 117
Bolsista: Jefferson da Silva Fernandes de Azevedo
Orientador: Pablo Javier Blanco

Gerência de Aplicações Científicas no Portal da Rede Nacional de Bioinformática (Bioinfo-Portal).....124
Bolsista: Mayconn Luiz Bispo dos Santos
Orientador: Kary Ann del Carmen Ocaña Gautherot, Antonio Tadeu Azevedo Gomes e Marcelo Monteiro Galheigo

Projeto de Iniciação Científica

Orientador: Prof. Fabio Porto[1]
Co-orientador: Prof. Douglas Ericson de Oliveira[1,2]
Aluno: André Muniz Demori[1,2]

[1]Laboratório Nacional de Computação Científica - LNCC
[2]Faculdade de Educação Tecnológica do Estado do Rio de Janeiro - FAETERJ

Indexação Métrica em Junções Espacias

Área do Conhecimento CNPq: 1.03.00.00-7 Ciência da Computação; 1.03.03.03-0 Banco de Dados
Período relatado: 01/08/2018 – 31/07/2019

Objetivos

Este projeto visa estudar, implementar e avaliar a adoção de índices métricos para auxílio na computação de junções espaciais, implementando nos softwares Apache Spark e SAVIME e técnicas de ML baseadas em métodos de indexação.

Introdução

Aplicações científicas envolvem o processamento de enormes conjuntos de dados. Na presença de dados espaciais, como na astronomia, os índices podem evitar varreduras desnecessárias de todo o conjunto de dados ao procurar objetos em uma vizinhança espacial. No entanto, diferentes estruturas de indexação existentes exibem desempenho variável. Neste trabalho, investigamos experimentalmente o desempenho de estruturas de indexação espacial considerando sua implementação em estruturas de Big Data, como o Spark. Nós privilegiamos, como critério de seleção, a quantidade de espaço de memória consumida e o tempo decorrido para responder às consultas de vizinhança. Consideramos três opções de indexação muito diferentes: Quad-tree, uma estrutura tradicional de indexação 2D; Slim-tree, um índice métrico e PH-tree, uma estrutura de indexação binária multidimensional. Mostramos que o Slim-tree aloca ordens de grandeza menos memória que seus concorrentes. Esse é um critério muito importante para estruturas na memória, como o Apache Spark. No entanto, à medida que a quantidade de objetos indexados aumenta, a árvore PH torna-se um vencedor claro em relação ao tempo decorrido da consulta.

Metodologia

Consideramos um contexto envolvendo grandes conjuntos de dados científicos de objetos espaciais. Cada localidade espacial do objeto é identificada por um valor em um sistema de coordenadas, como o sistema de coordenadas Equatorial 2D com base na ascensão reta (ra) e declinação (dec), com isso devemos usar esses dados e realizar a indexação para consultas espaciais.

Existe um número significativo de famílias de estruturas de indexação espacial. Nossa intenção neste trabalho não é fornecer uma avaliação extensa e completa deles, mas sim aplicar uma abordagem sistemática para comparar três importantes representantes dessas famílias.

Slim-tree:

Slim-tree um Método de Acesso Métrico (MAM). Esse método de acesso trabalha com objetos em um espaço métrico, composto de objetos set e uma função que estabelece uma distância entre quaisquer dois objetos do conjunto. Os MAMs selecionam objetos para se tornarem representantes de subconjuntos de dados. A partir daí, a distância de um novo elemento para cada representante c é calculada e organizada em sua estrutura de modo que o objeto se torne parte do subconjunto de dados do representante menos distante.

Em suma, a ideia básica de tais estruturas de dados é escolher um conjunto de elementos centrais e aplicar uma função de distância para agrupar os objetos remanescentes em subconjuntos apropriados (com relação a distâncias), Figura 1.

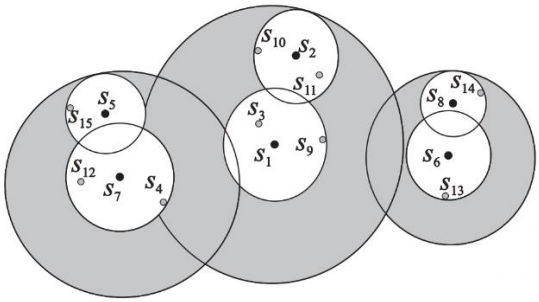


Figura 1 - Slim-tree – Distribuição Espacial

PH-tree:

PH-tree - PATRICIA-Hypercube-tree, Figura 2, pertence à família Quad-tree, mas fornece eficiência e escalabilidade muito melhores com maior suporte de dimensionalidade. É uma estrutura desequilibrada que basicamente se baseia na Quad-tree e divide o espaço em cada nó em várias dimensões. Isso reduz o número de nós na árvore, já que cada nó pode ter 2^k filhos, para k dimensões Ao mesmo tempo, a profundidade máxima da árvore é independente de k e igual ao número de bits no maior valor armazenado, ou seja, 8 ao armazenar valores de byte. Espera-se que a PH-tree seja muito bem dimensionada com grandes conjuntos de dados, em alguns casos, conjuntos maiores com 10^6 têm um desempenho melhor do que conjuntos de dados menores.

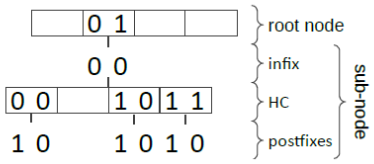


Figura 2 - Uma representação de uma árvore PH 2D com 3 entradas de 4 bits: (0001, 1000), (0011, 1000), (0011, 1010)

Quad-tree:

Quad-tree é uma estrutura de dados hierárquica bidimensional que recursivamente divide o espaço hierarquicamente em quadrantes. Quando objetos de alta dimensão, maiores que 2, devem ser indexados, outras estruturas da família Quad-tree oferecem propriedades de indexação semelhantes, como por exemplo o octree.

A abordagem Quad-tree produz árvores desequilibradas, que são sensíveis à distribuição de densidade espacial dos objetos que indexa. Um exemplo de representação desta estrutura de dados é mostrado na Figura 3.

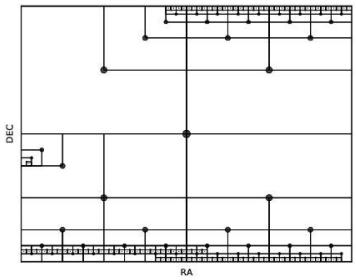


Figura 3 – Representação do particionamento recursivo da Quad-tree

O processo de indexação divide sucessivamente os objetos em um quadrante em quatro sub quadrantes. Os quadrantes geralmente são nomeados de acordo com sua direção em relação ao centro do nó pai, por exemplo: NW, NE, SW e SE.

Resultados e Discussão

No primeiro experimento, executamos uma consulta de intervalo que pesquisa 100 objetos no espaço de consulta e considera 100 distâncias, ambas geradas aleatoriamente. Assim, cada objeto procura vizinhos nas 100 distâncias. As consultas foram aplicadas a um conjunto de dados de 1.035.204 objetos do projeto SDSS SkyServer DR15. Os algoritmos para cada teste foram executados 30 vezes.

Para o segundo experimento, o objetivo foi analisar o desempenho das estruturas em conjuntos de dados com diferentes densidades. Para tanto, um ponto foi indexado no centro do espaço onde o conjunto de dados foi criado com limites definidos em relação à posição do ponto e da distância. Dez conjuntos de dados foram criados dentro desses limites. Os algoritmos para cada teste foram executados 10 vezes.

Resultados do primeiro experimento

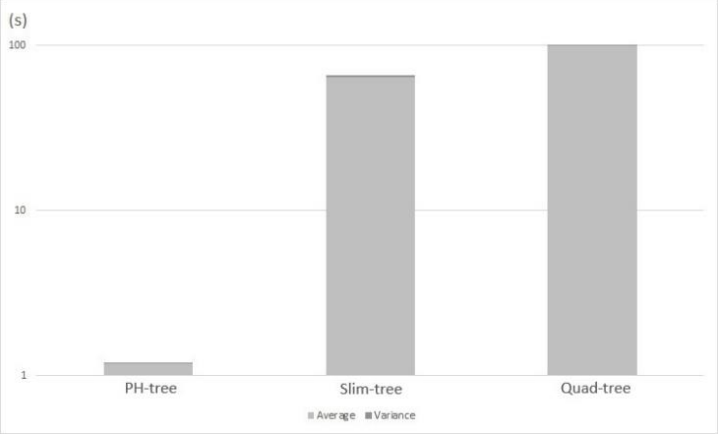


Figura 4 - Tempo de criação das estruturas

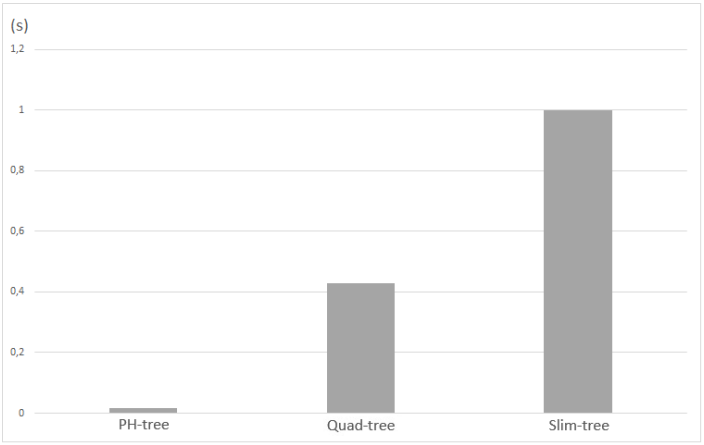


Figura 5 - Tempo de busca aos vizinhos – Slim-tree está para 1 e os outros valores são proporcionais

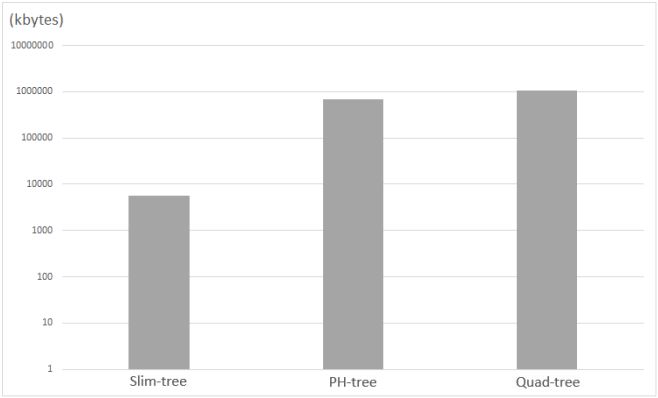


Figura 6 - Maximum Resident Set Size – escala logaritmica:

Resultados do segundo experimento com diferentes densidades:

Nos experimentos mostrados na seção no primeiro experimento usamos o algoritmo Minoccup, como uma opção para a estratégia Choosesubtree para tratar a inserção de objetos na Slim-tree. Esta opção permite, no momento da inserção, escolher o nó interno que possui a ocupação mínima entre os qualificados. Esta política gera árvores de menor altura, mas com um maior grau de sobreposição. Para os experimentos descritos nesta seção, decidimos usar a opção Mindist que escolhe o nó cujo representante tem a menor distância do novo elemento. Esta política gera árvores mais altas com um menor grau de sobreposição.

Figura 7 - Tempo de criação da árvore – escala logarítmica

Figura 8 - Tempo de busca aos vizinhos – escala logarítmica

Figura 9 - Maximum Resident Set Size – escala logaritmica:

A PH-tree apresentou melhores resultados em tempo de criação e tempo de busca. Sua estrutura binária se mostrou ser a mais eficiente.

Para consultar vizinhos sob diferentes densidades, Slim-tree e Quad-tree foram mais eficientes para uma pequena quantidade de objetos, mas mais uma vez o PH-tree mostrou-se mais eficiente à medida que a quantidade de dados aumentou. Isso pode estar associado ao mesmo problema levantado anteriormente, associado ao aumento de interseções de nós. Quanto às árvores quádruplas, as áreas densas aumentam a altura da árvore, levando à navegação recursiva por vários nós.

Conclusão

Para obter os resultados rodamos dois conjuntos de experimentos. Um considerou um conjunto de dados de astronomia real e um segundo conjunto usando um conjunto de dados sintético com densidade variável. Em ambos os casos, a estrutura do índice Slim-tree mostrou ordens de grandeza de tamanho de alocação de memória menor. Esse é um aspecto relevante para sistemas na memória que processam grandes conjuntos de dados. No entanto, para o tempo decorrido da consulta, a árvore PH mostra resultados mais consistentes. Sempre que o tamanho do conjunto de dados fosse de tamanhos mais realísticos, ele superaria as outras duas estruturas de dados. O Slim-tree foi eficiente para consultas sobre um pequeno número de objetos, mas perdeu o desempenho assim que o tamanho do conjunto de dados aumenta. O Quad-tree mostra resultados intermediários na alocação de memória e no tempo decorrido da consulta. Em vista desses resultados, deve-se avaliar esses dois eixos: (tempo decorrido da consulta de alocação de memória / intervalo) para escolher entre a árvore PH e a árvore quádrupla. Este projeto visa agora a implementação de Learned Indexes, onde a proposta é substituir as estruturas de indexação por modelos de ML. Este projeto foi tema de um artigo denominado *On The adoption of Spatial Indexing for Scientific Big Data*

Applications para o 34º SBBD. Esse relatório é um resumo do que foi abordado no artigo. Para mais detalhes, procurar os autores para a disponibilização do artigo.

Referências

Barioni, M. C. N., Razente, H. L., Traina, A. J., and Traina Jr, C. (2006). An efficient approach to scale up k-medoid based algorithms in large databases. In SBBD, pages 265–279.

Eldawy, A. and Mokbel, M. F. (2015). The ecosystem of spatialhadoop. SIGSPATIAL Special, 6(3):3–10.

Finkel, R. A. and Bentley, J. L. (1974). Quad trees a data structure for retrieval on composite keys. Acta informatica, 4(1):1–9.

Guttman, A. (1984). R-trees: a dynamic index structure for spatial searching. In Proceedings of the ACM International Conference on Management of Data - SIGMOD, pages 47–57. ACM.

Khatibi, A., Porto, F., Rittmeyer, J. G., Ogasawara, E., Valduriez, P., and Shasha, D. (2017). Pre-processing and indexing techniques for constellation queries in big data. In Proceedings of the International Conference on Big Data Analytics and Knowledge Discovery, pages 164–172. Springer International Publishing.

Ozsu, T. and Valduriez, P. (2011). Principles of Distributed Database Systems. Springer Verlag.

Porto, F., Khatibi, A., Rittmeyer, J. N., Ogasawara, E. S., Valduriez, P., and Shasha, D. E. (2018a). Constellation queries over big data. In XXXIII Simp´osio Brasileiro de Banco de Dados, SBBD 2018, Rio de Janeiro, RJ, Brazil, August 25-26, 2018., pages 85–96.

Porto, F., Rittmeyer, J., Ogasawara, E., Krone-Martins, A., Valduriez, P., and Shasha, D. (2018b). Point pattern search in big data. In Proceedings of the 30th International Conference on Scientific and Statistical Database Management, page 21. ACM.

Ramakrishnan, R., G. J. (2002). Database Management Systems. McGraw-Hill.
Samet, H. (1984). The quadtree and related hierarchical data structures. ACM Comput. Surv., 16(2):187–260.

Santos, L. F. D. (2012). Explorando variedade em consultas por similaridade. PhD thesis, Universidade de São Paulo.

Traina, C., Traina, A., Seeger, B., and Faloutsos, C. (2000). Slim-trees: High performance metric trees minimizing overlap between nodes. In International Conference on Extending Database Technology, pages 51–65. Springer.

You, S., Zhang, J., and Gruenwald, L. (2015). Large-scale spatial join query processing in cloud. In 31st IEEE International Conference on Data Engineering Workshops. IEEE.
Yu, J., Zhang, Z., and Sarwat, M. (2019). Spatial data management in apache spark: the geospark perspective and beyond. GeoInformatica, 23(1):37–68.

Zäschke, T., Zimmerli, C., and Norrie, M. C. (2014). The ph-tree: a space-efficient storage structure and multi-dimensional index. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data, pages 397–408. ACM.

Segurança e Privacidade em Redes de Anonimato

Bolsista: Daniel Veiga da Silva Antunes
Orientador: Fábio Borges de Oliveira
PIBIC - Programa Institucional de Bolsas de Iniciação Científica

Julho 2019

1 Objetivos

1.1 Objetivos Gerais

- Descrever as estruturas de algoritmos para manter o anonimato
- Estudar onde tem sido aplicado tais algoritmos
- Demonstrar o seu funcionamento matemático
- Realizar análise e comparação entre algoritmos simétricos e assimétricos

1.2 Objetivos Específicos

- Criação de uma rede de comunicação com sockets em linguagem C
- Descrever a técnica de Mix Network e o algoritmo do Tor
- Pesquisar quais os problemas em Mix Network
- Estudar e descrever DC-Net simétricas
- Estudar Teoria de Grupos
- Descrever as ideias de criptografia assimétrica em anonimato.

2 Introdução

A privacidade e segurança no mundo digital vem se mostrando a cada dia mais um desafio. Ao longo da história foram desenvolvidas diferentes ferramentas e técnicas visando o trafego seguro de informação no meio físico, surgindo assim as primeiras mensagens criptografadas. No meio digital o desafio ganha um novo patamar, a construção de uma rede de anonimato afim de manter a privacidade dos dados que ali circulam. Este trabalho visa estudar o funcionamento dos principais algoritmos de anonimato e suas aplicações, afim de criar uma rede

de comunicação segura em linguagem C através de sockets e dos algoritmos estudados.

3 Metodologia

Estudo Dirigido e Apresentação de Seminários.

4 Referências Bibliográficas

- William Stallings, Cryptography and Network Security, 2014, Prentice Hall
- Routo Terada, Segurança de Dados: Criptografia em Redes de Computadores, 2000, Edgard Blucher
- F. Borges, L. A. Martucci, M. Mühlhäuser, Analysis of Privacy-Enhancing Protocols Based on Anonymity Networks, 2012, IEEE Third International Conference on Smart Grid Communications (SmartGridComm)
- David Chaum, The Dining Cryptographers Problem: Unconditional Sender and Recipient Untraceability, 1988, Journal of Cryptology
- S. Shokranian, M. Soares, H. Godinho, Teoria dos números, 1999, Ed. UNB
- F. Borges, Privacy-preserving data aggregation in smart metering systems, 'Energy: From Smart Metering to the Smart Grid', Chap. 2, 2016, IET

Calibração de Modelos de Incrustação de $CaSO_4$ em Trocadores de Calor

Bolsista: Fernando Henrique Pereira Cardozo

Orientador: Prof^o Renato Simões

1 Objetivo

O principal objetivo do trabalho é implementar modelos computacionais que usem combinação de modelos existentes na literatura de forma a melhorar a precisão das previsões. Esse objetivo passa pelo estudo da calibração de modelos conhecidos da literatura. Através de uma série de dados para determinado modelo, tem-se por objetivo modelar de forma computacional uma previsão para condições do problema diferentes daquelas que se tem os dados.

2 Introdução

A incrustação de $CaSO_4$ em trocadores é um problema que afeta diversas industrias e para entender e prever sua ocorrência se faz uso de modelos matemáticos e numéricos.

Como o objetivo é encontrar o modelo com o melhor resultado e sabendo que tem-se inúmeros modelos na literatura é necessário que sejam feitas comparações entre os resultados encontrados por cada modelo e os dados utilizados por cada um deles para que se possa fazer uma comparação.

3 Material e Metodologia

No início desse trabalho foi estudo o método de Euler [1] para soluções numéricas. O Método aproxima as soluções através da derivada da função, como segue na equação 1.

$$y(t_i) = y(t_{i-1}) + h\frac{dy(t)}{dt}|_{t-t_{i-1}} \tag{1}$$

Através do paper *Fouling of Heat Transfer Surfaces - Matthias Bohnet* [2] foram retirados dados experimentais de modelos de incrustação para serem usados nos testes de previsão. A relação que rege o modelo de Bohnet é apresentada nas equações 2, 3 e 4.

$$\frac{dm_f}{dt} = \dot{m}_d - \dot{m}_r \tag{2}$$

$$\left(\frac{dm_f}{dt}\right)_d = \dot{m}_d = \beta\left\{\frac{1}{2}\left(\frac{\beta}{k_r}\right) + (C_F - C_s) - \sqrt{\frac{1}{4}\left(\frac{\beta}{k_r}\right)^2 + \left(\frac{\beta}{k_r}\right)(C_F - C_s)}\right\} \quad (3)$$

$$\left(\frac{dm_f}{dt}\right)_r = \dot{m}_r = \frac{K_6}{P}\rho_f(1 + \delta\Delta T)d_p(\rho^2\eta g)^{1/3}\frac{m_f}{\rho_f}w^2 \quad (4)$$

Para melhorar a precisão dos resultados encontrados, é feita uma calibração do coeficiente de transferência de massa (β). Para fazer essa calibração, foram utilizados dois métodos computacionais, Luus-Jaakola e PSO.

3.1 Luus-Jaakola

O método de Luus-Jaakola [3] consiste num modelo iterátivo, que busca convergir a solução através de uma solução qualquer, dita solução atual, e de um conjunto de pontos aleatórios gerados dentro do domínio de busca. Cada um dos pontos aleatórios é testado na função que deseja-se minimizar e o valor comparado ao valor da solução atual. Caso algum dos pontos mostre um desempenho melhor, seu valor da função torna-se a solução atual. Ao final da iteração, o domínio de busca é reduzido de um fator de contração, gerando novos aleatórios dentro do novo domínio e repete-se o processo até que a condição de parada seja atingida.
Um pseudo-cógido do algoritmo é apresentado na figura 1.

```
Define-se a solução atual f(atual)
Enquanto condição de parada não atingida:
        Inicia-se o vetor aleatório X dentro do domínio de busca
        Para cada i em X:
                Calcular f(i)
                Se f(i) melhor que f(atual)
                        f(atual) = f(i)
        Reduzir domínio de busca
```

Figura 1: Pseudo-código do Algoritmo LJ.

3.2 PSO

O PSO (*Particle Swarm Optimization*) [4] é um algoritmo populacional que foi desenvolvido baseado na movimentação de bandos de pássaros e cardumes de peixes em busca de alimento. O algoritmo consiste na geração de uma população aleatória, que representa os individuos.

Cada partícula é representada por um vetor velocidade e um vetor posição, sendo o vetor velocidade caracterizado pelo balanço entre a busca local (conhecimento do melhor local já visitado pela partícula - *pbest*) e a busca global (conhecimento do melhor local já

visitado pela melhor partícula do grupo - *gbest*). O vetor velocidade, após sua atualização, corrige o vetor posição da partícula.

Cada partícula do grupo é avaliada a cada geração por uma função *fitness*, que avalia o quão boa é a posição encontrada por esse indivíduo em relação à solução do problema.

A cada interação o programa atualiza a posição das partículas, recalcula sua função *fitness*, descobre quem é a melhor partícula da nova população e para cada partícula passa a informação de qual a melhor posição na qual aquela partícula já esteve, desde que foi gerada, e qual a melhor posição já visitada pela melhor partícula do grupo.

Um pseudo-código do PSO é apresentado na figura 2.

Inicialização

Para cada partícula ***i*** *na população:*

Para cada dimensão ***d****:*

Gerar $\boldsymbol{x}_{i,d}$ *aleatório*

Guardar no vetor $\boldsymbol{X}_i$

Gerar $\boldsymbol{v}_{i,d}$ *aleatório*

Guardar no vetor $\boldsymbol{V}_i$

Avaliar a fitness de $\boldsymbol{X}_i$

Gerar ***fbest*** *como cópia de* $\boldsymbol{X}_i$ *adicionando os valores das fitness a cada vetor.*

Salvar em ***gbest*** *a posição que possui a melhor fitness e o valor dessa fitness.*

Loop

Repetir até o critério de parada ser atingido

Para cada partícula ***i****:*

$V_i(k+1) = w.X_i(k) + R_1.C_1.(P_i(k) - X_i(k)) + R_2.C_2.(P_G(k) - X_i(k))$

$X_i(k+1) = V_i(k+1) + X_i(k)$

Avaliar a fitness do novo $\boldsymbol{X}_i$

Se fitness(novo $\boldsymbol{X}_i$*) >* ***gbest:***

gbest *= fitness(novo* $\boldsymbol{X}_i$*)*

Figura 2: Pseudo-código do Algoritmo PSO.

4 Resultados

A tabela 1 mostra os resultados dos coeficientes de transferência de massa (β) otimizados, tanto pelo algoritmo Luus-Jaakola quanto pelo PSO.

Velocidade (m/s)	Beta (β) PSO	Erro	Beta (β) LJ	Erro
0.50	$7.62429 \cdot 10^{-5}$	$3.80524 \cdot 10^{-5}$	$7.60840 \cdot 10^{-5}$	$3.81199 \cdot 10^{-5}$
0.55	$4.13708 \cdot 10^{-5}$	$1.93000 \cdot 10^{-5}$	$4.13483 \cdot 10^{-5}$	$1.93039 \cdot 10^{-5}$
0.60	$4.14161 \cdot 10^{-5}$	$1.56443 \cdot 10^{-5}$	$4.14091 \cdot 10^{-5}$	$1.56440 \cdot 10^{-5}$

Tabela 1: Valores de Beta Otimizados.

Pode-se observar na tabela 1 que os valores para ambos os métodos de otimização se equiparam, mostrando que ambos são boas formas de se trabalhar com otimização, assim como apresentando resultados de forma precisa.

No gráfico apresentado na figura 3 são mostrados os dados experimentais para 3 diferentes velocidades e as curvas encontradas utilizando-se o método de Euler, equações 1, 2, 3 e 4 para os valores de coeficiente de transferência de massa (β) otimizados pelo algoritmo PSO. A escolha dos valores encontrados pelo PSO se dá por terem sido encontrados valores com menor erro.

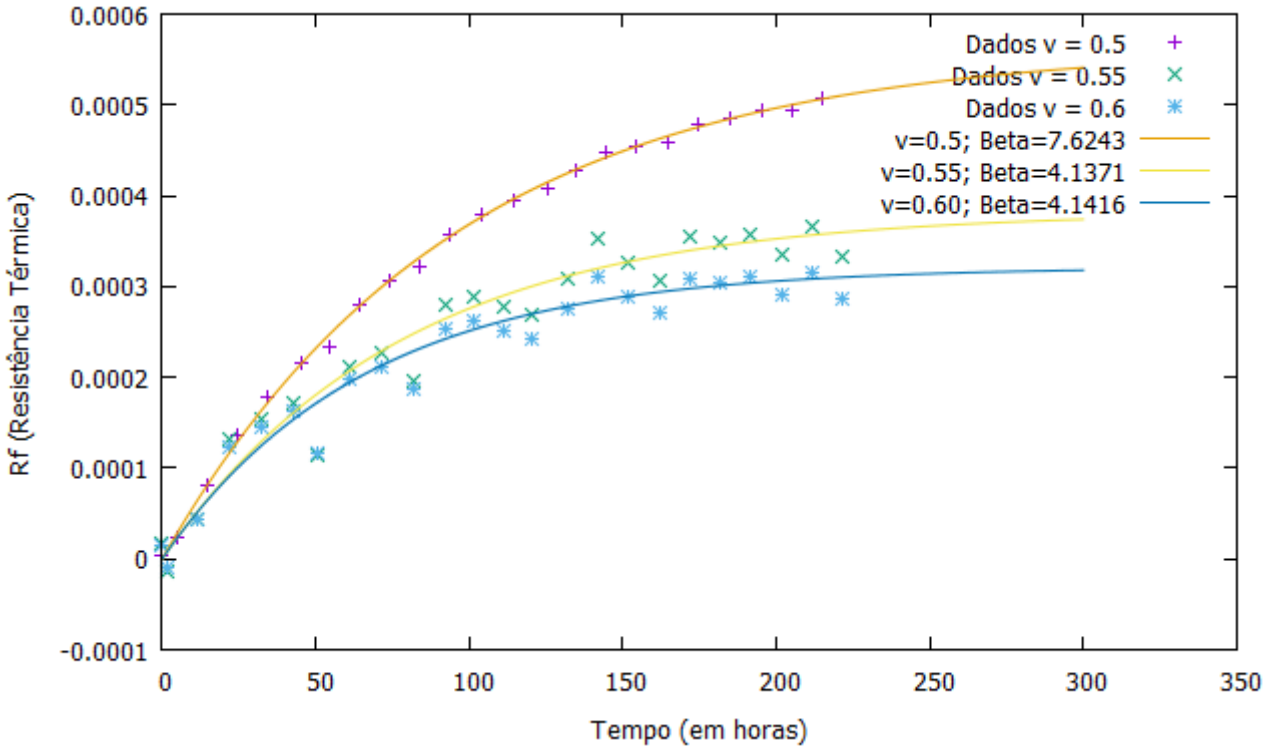


Figura 3: Gráfico Tempo x Rf para diferentes velocidades

5 Conclusões

A partir dos resultados apresentados na seção 4 pode-se observar que obteve-se uma boa otimização do parâmetro. Partindo dessa otimização objetiva-se criar uma correlação empírica entre a velocidade do escoamento e o coeficiente de transferência de massa, de forma a sabendo-se a velocidade do escoamento, obter-se o coeficiente de transferência de massa com uma margem de erro aceitável.

Como o coeficiente de transferência de massa (β) está intrinsecamente ligado ao fator de incrustação (R_f), que representa a resistência térmica da camada de incrustação, apresentado no gráfico 3, se conseguir-se uma correlação que estime o valor (β) através da velocidade de escoamento, será possível fazer previsões dos fatores de incrustação a partir do valor de velocidade do escoamento, que é um parâmetro de fácil medição.

Referências

[1] Gerald RECKTENWALD. Numerical integration of ordinary differential equations for initial value problems. 2006.

[2] Matthias BOHNET. Fouling of heat transfer surfaces. 1987.

[3] Antônio José da Silva Neto; José Carlos Becceneri e Haroldo Fraga de Campos Velho. *Inteligência Computacional Aplicada a Problemas Inversos em Trasnferência Radiativa*. EdUERJ, 2006.

[4] Eberhart R. C. e Kennedy J. A new optimizer using particle swarm theory. 1995.

RELATÓRIO DE ATIVIDADES

Título do Projeto: Análise de Dados de Proveniência em Aplicações de Biologia Computacional.

Nome do bolsista: Guilherme da Silva Vieira

Nome do orientador:

D.Sc. Kary Ann del Carmen Ocaña Gautherot (Tecnologista Pleno – LABINFO/LNCC)

Nome do Coorientador:

D.Sc. Fábio André Machado Porto (Tecnologista Sênior – DEXL/LNCC)

D.Sc. Douglas Ericson Marcelino de Oliveira (Pós-Doutor – DEXL/LNCC)

Tipo de bolsa: PIBIC

Período do relatório: 01/01/2019 a 12/07/2019

INTRODUÇÃO

O Projeto de Iniciação Científica (IC) se desenvolve no nível de estágio supervisionado sob coordenação da orientadora Kary Ocaña (LNCC) e com bolsa IC financiada pelo CNPq. O mesmo se enquadra nas pesquisas relacionadas aos projetos institucionais do LNCC, especificamente da Rede Nacional de Bioinformática (RNBio) e do Portal-Bioinfo (https://bioinfo.lncc.br/). O Portal-Bioinfo visa a execução em larga escala de aplicações de bioinformática usando recursos computacionais paralelos e distribuídos a fim de diminuir o grande tempo de processamento das execuções e no apoio às pesquisas da comunidade científica de bioinformática.

O Projeto de IC visa o desenvolvimento de um banco de dados de proveniência para o Portal-Bioinfo. Assim o projeto apresenta as seguintes etapas: (1) revisar a bibliografia sobre banco de dados, Sistemas de Gerência de Banco de Dados (SGBD), *gateways* científicos, bioinformática e da infraestrutura do supercomputador Santos Dumont (SDumont, https://sdumont.lncc.br); (2) revisar tecnologias de SGBD e Web; (3) levantar uma modelagem conceitual do projeto de banco de dados do Portal-Bioinfo; (4) levantar um sistema de controle e gerência do Portal- Bioinfo no ambiente do Sistema Nacional de Processamento de Alto Desempenho (SINAPAD, https://www.lncc.br/sinapad/) e (5) acoplar o uso de PostgreSQL para a criação do esquema e para levantar consultas no Portal-Bioinfo.

O desenvolvimento do banco de dados para o Portal-Bioinfo irá prover um banco de dados que poderá ser analisada e consultada, que é suportada pelo acesso aos dados de proveniência, e que visará otimizar a funcionalidade do portal.

OBJETIVOS

Os objetivos do Projeto de IC são: (1) implementar um modelo de Banco de Dados conceitual no apoio ao Portal-Bioinfo e suportado por tecnologias de SGBD, acoplado aos recursos computacionais do SINAPAD e da infraestrutura do supercomputador Santos Dumont, (2) implementar funcionalidades de gerência dos dados de proveniência gerados pelas aplicações do portal por meio de consultas no banco de dados e (3) aplicar testes de desempenho para validar a consistência do banco de dados.

METODOLOGIA

A primeira etapa envolveu uma revisão bibliográfica sobre o PostgreSQL v.11 utilizado como sistema de gerência de banco de dados (SGBD); e sobre o pgAdmin4 v4.3 como ferramenta para acessar o banco de dados.

Na segunda etapa analisamos o problema proposto a fim de solucioná-lo, montando o modelo conceitual para melhor visualização do projeto proposto.

RESULTADOS E DISCUSSÃO

Sobre o desenvolvimento do banco de dados, primeiramente foi necessário o levantamento bibliográfico sobre a gerência e modelos de banco de dados, com foco no estudo e implementação da entidade-relacionamento. A Figura 1 a seguir mostra um fragmento do modelo total desenvolvido e as respectivas entidades, que foram acopladas e desenvolvidas no presente Projeto de IC. A entidade *Arquivo* se refere a entrada que será consumida por uma aplicação e relacionada a entidade Execução. A Figura 2 apresenta todas as tabelas que foram desenvolvidas para armazenar os dados de proveniência do portal

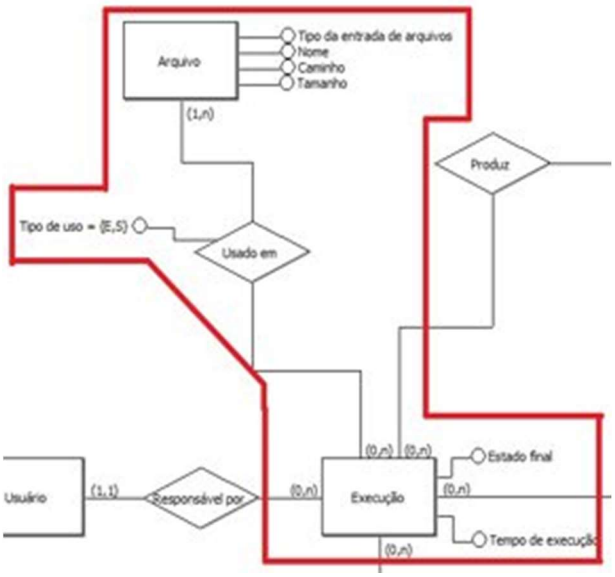


Figura 1. Modelo entidade-relacionamento desenvolvido para o Portal-Bioinfo.

Estudo de Caso:

- Mapeamento do modelo conceitual ER: Arquivo e Execução são as únicas entidades existentes no Portal-Bioinfo (Figura 1)

Arquivo (tipo, nomeArquivo, caminho, tamanho); Execução (estadoFinal, tempo de execução); Usado em (nomeArquivo, estadoFinal, tipo de uso)

- Arquivo é responsável por armazenar as informações do dado de entrada, tais como: tipo de entrada, nome e tamanho do arquivo e o local de armazenamento do mesmo
- Execução armazena informações de tempo de execução (início e fim) e estado da execução (em espera, executando, finalizado com erro e finalizado sem erro)
- Usado é o relacionamento entre as entidades Arquivo e Execução

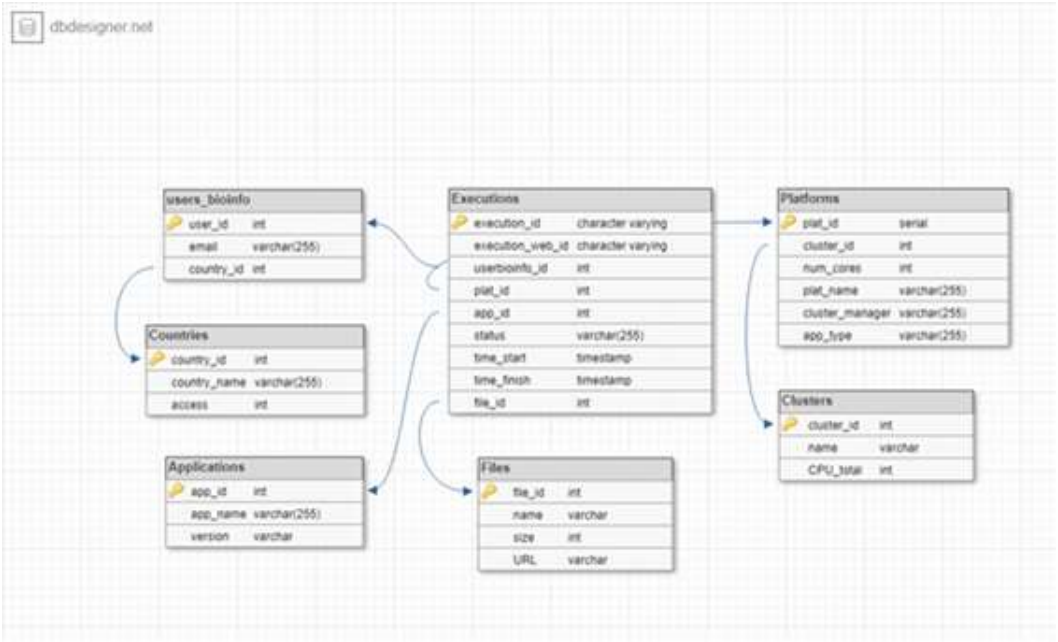


Figura 2. Tabelas do modelo desenvolvidas para o Portal-Bioinfo

O banco de dados está sendo montado tendo como base as tabelas da Figura 2, criadas na plataforma online dbdesigner. Nesse desenvolvimento estão sendo criadas as tabelas da Figura 2, seus relacionamentos, assim como também funções de preenchimento utilizados para facilitar inserções, como nas consultas a seguir com comandos de inserção na tabela *executions*:

```
CREATE OR REPLACE FUNCTION
insert_execution(execution_id varchar, execution_webid varchar, app_id int,
file_id int, plat_id int, status varchar, time_start time, time_finish time)
      RETURNS text  AS $$
      BEGIN
              INSERT INTO executions (execution_id, execution_webid, app_id,
      file_id, plat_id, status, time_start, time_finish) VALUES (execution_id,
      user_bioinfo_id, cluster_id, app_id, file_id, status, time_start,
      time_finish);
          RETURN 'Execução cadastrada com sucesso';
      END;
      $$ LANGUAGE plpgsql;
```

```
Para utilizar a função:
SELECT insert_execution('a','b',1,1,1,'', '22:30:00', '22:32:47');
```

Onde os valores dentro dos parênteses ainda são apenas para teste, preenchendo os atributos execution_id(chave primária), execution_webid(id proveniente do portal), app_id(chave estrangeira), file_id(chave estrangeira), plat_id(chave estrangeira) status(estado da execução), time_start(hora de início da execução), time_finish(hora do fim da execução).

CONCLUSÕES

Apresentamos resultados sobre o desenvolvimento e implementação de um banco de dados acoplado à infraestrutura do Portal-Bioinfo, que permitirá armazenar dados de proveniência, por exemplo, das execuções de todas as aplicações do portal. Nesse projeto salientamos que esse modelo de entidade-relacionamento para o Portal-Bioinfo foi implementado, o que até então não existia. Dessa maneira, o desenvolvimento e implementação de um modelo conceitual ER para o Portal-Bioinfo permitirá a garantir a consistência na gerência e acesso aos dados científicos, de proveniência e de desempenho.

Esse modelo irá sustentar o acesso aos dados para estudos realizados em paralelo no Bioinfo-Portal, análises de predição via aprendizado de máquina e o desenvolvimento da interface Web inteligente.

REFERÊNCIAS

Documentação PostgreSQL: https://www.postgresql.org/docs/current/ddl-basics.html Acesso em: Abr/Mai/Jun/Jul 2019(disponível online)

Dbdesigner: https://www.dbdesigner.net/ Acesso em Mai/Jun/Jul 2019 (disponível online)

RELATÓRIO DE ATIVIDADES

Título do Projeto: Desenvolvimento de Módulos Paralelo-Híbrido de Bioinformática para Ambientes GPU de Supercomputação

Nome do bolsista: Guilherme Freire da Silva Dornelas

Nome do orientador:

D.Sc. Carla Osthoff Ferreira de Barros (Tecnologista Sênior – CENAPAD/LNCC)

Nome do coorientador:

D.Sc. Kary Ann del Carmen Ocaña Gautherot (Tecnologista Pleno – LABINFO/LNCC)

D.Sc. André Elias Rodrigues Soares (Pós-Doutor – LABINFO/LNCC)

Tipo de bolsa: PIBIC

Período do relatório: 01/01/2019 a 31/07/2019

OBJETIVOS

O objetivo principal do projeto visa adaptar algoritmos de bioinformática em ambientes de PAD, de código otimizado em GPU, para utilizar de forma eficiente a GPU K40 do SDumont que possui uma escala muito maior de processadores e memória local. Fazendo uso dessa ferramenta de processamento gráfico iremos solucionar problemas envolvendo a demora no cálculo de probabilidades nas árvores, gargalos e melhorar o desempenho das aplicações do portal BioInfo, que estão instaladas no SDumont.

O objetivo secundário era realizar estudos de desempenho e escalabilidade para identificar o uso mais eficiente de recursos computacionais a serem alocados, segundo parâmetros de entrada de cada algoritmo e características dos dados de entrada como tamanho, tipo, localização e contribuir com as pesquisas multidisciplinares no LNCC. Integrando assim as pesquisas desenvolvidas pelo Centro Nacional de Processamento de Alto Desempenho (CENAPAD), o Sistema Nacional de Processamento de Alto Desempenho (SINAPAD) e o Laboratório de Bioinformática (LABINFO).

INTRODUÇÃO

A Computação de Alto Desempenho (CAD) é um termo ligado ao uso de supercomputadores ou clusters que executam atividades com uma grande demanda de processamento e.g. cálculos numéricos complexos. Como forma de colaborar nesse processo tem sido aplicado o uso de GPU (Unidade de Processamento Gráfico), para tornar a execução mais dinâmica.

O presente projeto trabalha em conjunto com o SINAPAD e o LABINFO (Laboratório de Bio Informática) do LNCC que faz o uso de várias aplicações que são utilizadas pelos pesquisadores, como o Beast2, Mr. Bayes, entre outras ferramentas de perfilamento, as quais estão instaladas no SDumont. Os materiais de pesquisas do laboratório que necessitam de processamento passam por esses programas, todo esse processo demanda muito tempo de execução.

Atualmente, existe um interesse crescente pela comunidade científica de bioinformática em explorar GPU com o intuito de usufruir de um incremento no poder computacional. A questão importante é orquestrar tarefas em um conjunto distribuído de recursos com custos minimizados, garantindo integridade e consistência nas informações. O SDumont permitirá a alocação dessas tarefas, ele conta com 198 Nós de computação B715 (thin node) com GPUs K40.

METODOLOGIA

A presente proposta visa análise e desempenho de escalabilidade em ambientes de GPU no Santos Dumont, explorando linguagens como OpenCL e NVidia (CUDA).

O problema será atacado em três frentes:

I. Migrar os programas multi-plataforma para análise filogenômicas de sequências moleculares Beast, Beast2, MrBayes e ExaML para o ambiente do SDumont;

II. Realizar análises de desempenho e escalabilidade em ambientes GPU no Sdumont, i.e., tempo, speedup, eficiência, memória e CPU consumidos;

III. Desenvolver novos módulos em MPI para esses programas que permitam a comunicação de forma paralela e distribuída em múltiplas GPU.

O Beast2 é o primeiro a ser migrado para a GPU no SDumont. Ele é um programa multi-plataforma para análise Bayesiana que usa MCMC para calcular o espaço da árvore ponderada à sua probabilidade posterior. A maior parte do tempo computacional é gasto no cálculo de probabilidades nas árvores. Para contornear esse problema, o Beast2 usa a biblioteca BEAGLE que calcula eficientemente a verossimilhança em CPU e GPU.

RESULTADOS E DISCUSSÃO

A proposta inicial foi desenvolver pesquisas para entender cada parte do projeto, alguns artigos foram utilizados como: *Avaliação do RAxML no Supercomputador Santos Dumont,* para verificar cada versão utilizada e análise comparativa sobre o impacto produzido pela variabilidade nas configurações de ambiente, programas e características dos dados genômicos;

Artigo Simulação Monte Carlo em algoritmos MCMC: Estudo Comparativo e Simulações, **MCMC:** Compreendem uma classe de algoritmos para amostragem a partir de uma distribuição de probabilidade. Muito utilizado em estatística bayesiana, física computacional, biologia computacional e linguística computacional, está conectado ao projeto pelo programa MrBayes;

Artigo BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics, a Beagle é uma biblioteca de alto desempenho, pode fazer uso de processadores altamente paralelos, como aqueles em placas gráficas (GPUs), a mesma trabalha em conjunto com o *Beast e Beast2, fazendo a alternância entre os cáculos de verossimilhança em CPU e GPU.*

Os principais programas que fazem parte do projeto: *Beast, Beast2, MrBayes e ExaML* e as ferramentas de perfilagem do SDumont, foram exploradas, buscando o entendimento de cada uma, suas variações com CPU e GPU, para assim encontrar o melhor resultado entre elas; e conteúdo voltado ao funcionamento e configurações do SDumont para melhor entendimento de todas as especificações técnicas.

A etapa seguinte estava relacionada a compreensão do funcionamento de uma API que é um conjunto de rotinas e padrões de programação para acesso a um aplicativo de software, estudos de algoritmos de

estatística bayesiana que está ligado as aplicações do projeto, o uso de módulo, de montadores de sequencias, ferramentas para estimativa de máxima verossimilhança.

O funcionamento da biblioteca de alto desempenho BEAGLE que pode executar os cálculos centrais no coração da maioria dos pacotes de filogenética Bayesiana. A biblioteca é aplicada junto ao Beast que é multi-plataforma para análise Bayesiana de sequências moleculares usando MCMC, para calcular a média do espaço da árvore, de modo que cada árvore é ponderada proporcionalmente à sua probabilidade posterior. O Beast2 usa a biblioteca BEAGLE que calcula eficientemente a verossimilhança em CPU e GPU.

Algumas pesquisas foram efetuadas sobre a filogenética com o intuito de compreender os dados gerados pelos programas. Os programas são executados via linha de comando do terminal Linux, com isso foram estudados alguns comandos básicos e outros para melhor performance das aplicações e fácil utilização dos mesmos, os quais estão diretamente ligados a execução, modificação dos códigos e visualização de tempo.

Este projeto propõe dar continuidade às pesquisas em andamento de otimização nas execuções de algoritmos de filogenômica no SDumont, por meio de uma solução baseada em API e na construção de módulos voltados para execução em GPU de diferentes fabricantes. A solução permitirá a execução em ambientes heterogêneos visando ganhos de desempenho para retornar o resultado da execução e comparação em tempo adequado.

CONCLUSÕES

Durante a fase de projeto alguns desafios inerentes ao processo foram enfrentados, como encontrar cada módulo que deveria integrar ao SDumont (módulo é um artefato de programação (código) que pode ser desenvolvido e compilado separadamente de outras partes do programa), para o melhor emprego dos programas. A busca por definir qual a variação entre CPU e GPU foi algo bem minucioso, pois através dela iriamos receber os resultados, que são primordiais, envolvendo a duração de cada processo. Atualmente os resultados de desempenho explorando o Beast no SDumont estão sendo analisados.

REFERÊNCIAS BIBLIOGRÁFICAS

• K. Ocaña, J. Pedro, M. Coelho, M. Galheigo, e C. Osthoff, “Avaliação do uso eficiente de algoritmos paralelos de filogenia em supercomputadores,” in Anais do 12o BreSci -

Brazilian eScience Workshop, Natal, RN, 2018.

• C.-L. Hung, Y.-S. Lin, C.-Y. Lin, Y.-C. Chung, e Y.-F. Chung, “CUDA ClustalW: An efficient parallel algorithm for progressive multiple sequence alignment on Multi-GPUs,” Comput Biol

Chem, vol. 58, pp. 62– 68, May 2015.

• Z. Yin, H. Lan, G. Tan, M. Lu, A. V. Vasilakos, e W. Liu, “Computing Platforms for Big Biological Data Analytics: Perspectives and Challenges,” Computational and Structural

Biotechnology Journal, vol. 15, pp. 403–411, 2017.

• Z. Jin e J. D. Bakos, “Extending the BEAGLE library to a multi-FPGA platform,” BMC Bioinformatics, vol. 14, no. 1, p. 25, 2013.

• J. P. Huelsenbeck e F. Ronquist, “MRBAYES: Bayesian inference of phylogenetic trees,” Bioinformatics, vol. 17, no. 8, pp. 754–755, Aug. 2001.

RELATÓRIO DE ATIVIDADES

Título do Projeto: Avaliação de métodos de aprendizado de máquina em aplicações de Bioinformática.

Nome do bolsista: Isabela Canuto Ramos

Nome do orientador:

D.Sc. Kary Ann del Carmen Ocaña Gautherot (Tecnologista Pleno – LABINFO/LNCC)

Nome do Coorientador:

D.Sc. Fábio André Machado Porto (Tecnologista Sênior – DEXL/LNCC)

D.Sc. Douglas Ericson Marcelino de Oliveira (Pós-Doutor – DEXL/LNCC)

Tipo de bolsa: PIBIC

Período do relatório: 01/01/2019 a 12/07/2019

INTRODUÇÃO

O Projeto de Iniciação Científica (IC) se desenvolve no nível de estágio supervisionado sob coordenação da orientadora Kary Ocaña (LNCC) e com bolsa IC financiada pelo CNPq. O mesmo se enquadra nas pesquisas relacionadas aos projetos institucionais do LNCC, especificamente da Rede Nacional de Bioinformática (RNBio) e do Portal-Bioinfo (https://bioinfo.lncc.br/). O Portal-Bioinfo visa a execução em larga escala de aplicações de bioinformática usando recursos computacionais paralelos e distribuídos a fim de diminuir o grande tempo de processamento das execuções.

O Projeto de IC visa o desenvolvimento de uma versão do Portal-Bioinfo mais segura, eficiente e interativa. Dessa maneira, o projeto apresenta as seguintes atividades: (1) mapear o modelo do banco de dados e a estrutura do Bioinfo-Portal, extrair os dados de proveniência e desempenho por meio de consultas a base de dados MySQL e PostgreSQL; (2) organizar e formatar os dados obtidos no formato (*.csv) e aplicar técnicas de decisão de árvore para a construção de modelos preditivos usando o programa Orange e sci kit learn e (3) organizar e formatar os dados obtidos (*.csv) e aplicar técnicas de decisão de árvore para a construção de modelos preditivos usando os programas Orange Data Mining e Scikit Learn. A criação de modelos de predição, proporcionará um uso mais eficiente dos recursos computacionais e paralelos do supercomputador Santos Dumont (SDumont), pois serão usadas as predições para otimizar a configuração de uso dos clusters (tipo e quantidade) tal que as execuções alcancem uma eficiência maior (usando como cut-off 75%) o que permitirá melhor disposição e uso, quando alocadas as máquinas para um determinado experimento.

OBJETIVOS

O principal objetivo do Projeto de IC é implementar uma nova versão do Portal-Bioinfo (Portal de Bioinformática do LNCC) v2.0 mais segura e eficiente, adaptada e acoplada para ser

executada no Santos Dumont. Nesse trabalho em específico, essa nova versão aborda abordagens de pesquisa para obter alocações de máquinas eficientes (o melhor possível) na execução de experimentos científicos no supercomputador. Esse objetivo principal se baseia em outros objetivos específicos do projeto como são explorar técnicas de aprendizado de máquina e mineração de dados e implementar um módulo para análise de dados do portal que possibilite a criação de modelos de predição e classificação. Dessa maneira, a escolha do número e tipo de clusters no SDumont será baseada na melhor decisão, ouseja aquelas alocações que gerem no mínimo 75% de eficiência no uso dos recursos computacionais.

ATIVIDADES DESENVOLVIDAS

A primeira etapa consistiu revisar a bibliografia sobre aplicações, grade, workflows, bioinformática e SDumont. Em seguida foi estudado o programa Orange data mining e desenvolvida uma árvore de decisão baseado nos dados de proveniência e execução do programa de bioinformática RAxML, relacionados a execuções no SDumont. Esse mesmo experimento usado pelo Orange foi replicado com o programa scikit-learn, com o objetivo de comparar os resultados obtidos em cada software.

Ao longo deste período foram realizadas as seguintes atividades: construção de um modelo de predição com as ferramentas Orange data mining (status finalizado) e scikit learn (status em andamento). Os resultados obtidos com o uso do programa Orange, foram satisfatórios, foi gerada uma árvore de decisão, que utilizou dados de jobs executados no supercomputador Santos Dumont, partindo do critério “class” (usado como target na classificação), que foi dividido em três categorias (baixo, alto, médio). Sendo assim, foi determinado como ponto de corte um cut-off de 75% para determinar o uso otimizado (parâmetros a serem levados em consideração) no uso dos clusters do SDumont.

O programa scikit learn, foi utilizado e demonstrou também bons resultados, os que ainda estão em aprimoramento e sendo analisados, através de um script desenvolvido na linguagem python. O scikit learn utilizou a mesma base de dados e target usados no programa Orange. Como resultados preliminares, a árvore gerada demonstrou apenas resultados satisfatórios com target classificado como baixo, o que ainda não é ideal, sendo assim esse estudo ainda está em andamento e os resultados analisados para serem confrontados e comparados com o Orange.

RESULTADOS E DISCUSSÃO

O Orange data mining, é um programa de mineração de dados e aprendizado de máquina. Os dados usados para gerar a árvore de decisão, se encontram presentes em um arquivo em formato csv (Figura 1), onde as vírgulas separam as colunas, que representam os dados de proveniência e execuções do experimento. A última coluna “tag” representam os três rótulos usados para classificar os valores de eficiência em low, medium, high.

Os dados apresentavam um desbalanceamento, pois 80% pertenciam ao rótulo low pelo que foi necessário discretizar os dados com uso do próprio Orange (Figura 2). Dessa maneira foi obtida uma tabela mais equilibrada o que permitiu gerar a árvore de decisão (Figura 3).

```
environment , workflow ,  datasize  , bootstrap , class ,tag
sd_cpu,raxml_hibrid,3.2,2000,30,low
sd_cpu,raxml_hibrid,3.2,2000,30,low
sd_cpu,raxml_hibrid,3.2,1000,40,low
sd_cpu,raxml_hibrid,3.2,1000,40,low
sd_cpu,raxml_hibrid,3.2,100,50,low
sd_cpu,raxml_hibrid,3.2,100,50,low
sd_cpu,raxml_hibrid,3.2,2000,60,low
sd_cpu,raxml_hibrid,3.2,2000,60,low
sd_cpu,raxml_hibrid,3.2,1000,70,low
sd_cpu,raxml_hibrid,3.2,1000,70,low
sd_cpu,raxml_hibrid,3.2,1000,70,low
sd_cpu,raxml_hibrid,3.2,1000,70,low
sd_cpu,raxml_hibrid,79,100,80,low
sd_cpu,raxml_hibrid,79,100,80,low
sd_cpu,raxml_hibrid,79,100,80,low
sd_cpu,raxml_hibrid,79,100,80,low
sd_cpu,raxml_hibrid,792,100,80,low
sd_cpu,raxml_hibrid,21000,1000,80,low
sd_cpu,raxml_hibrid,3.2,2000,80,low
sd_cpu,raxml_hibrid,79,2000,80,low
sd_cpu,raxml_hibrid,79,2000,80,low
sd_cpu,raxml_hibrid,79,100,80,low
sd_cpu,raxml_hibrid,79,100,80,low
sd_cpu,raxml_hibrid,79,100,80,low
sd_cpu,raxml_hibrid,79,100,80,low
```

Figura 1. Arquivo em formato de entrada “csv” usado no aprendizado de máquina

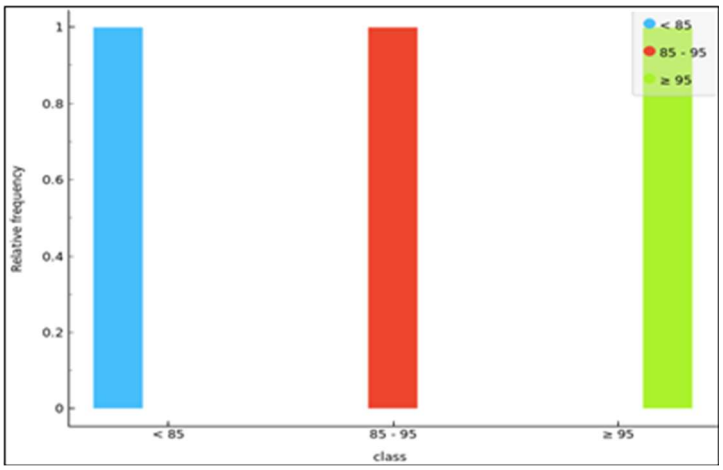


Figura 2. Distribuição dos dados nas classes low, medium, high

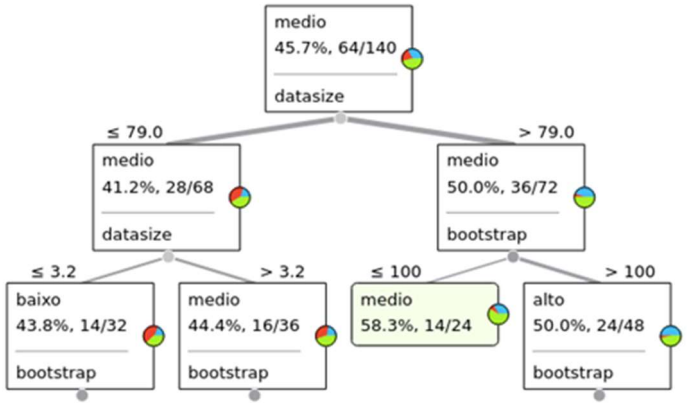


Figura 3. Árvore de decisão obtida pelo Orange

Um outro programa usado na técnica árvore de decisão foi o scikit-Learn. A **Erro! Fonte de referência não encontrada.** apresenta o script em python desenvolvido para gerar a árvore de decisão, a qual está em fase de análise.

```
#importação da biblioteca python: análise de dados e estruturas de dados.
import pandas as pd

#importação da biblioteca tree, responsável pela construção da árvore.
```

```
from sklearn import tree

#linha que declara as colunas presentes no arquivo csv usado como base de dados.
nomes_colunas  = ['datasize','bootstrap','classificador']

#linha que realiza a leitura do arquivo csv
leitura = pd.read_csv("teste2.csv", names=nomes_colunas)

#colunas_recursos declara colunas que serão usadas para gerar a árvore de decisão
colunas_recursos = ['datasize','bootstrap']

#X declara variável que será chamada para leitura e representa colunas de recurso.
X = leitura[colunas_recursos]

#y declara variável que será chamada para leitura e representa a coluna
classificadora, usada como target para a árvore.
y = leitura.classificador

#clf é o estimador de instância, e tree.DecisionTreeClassifier() é a saída
múltipla da árvore de decisão.
clf = tree.DecisionTreeClassifier()

#o comando fit é responsável por treinar o modelo.
clf = clf.fit(X,y)

#plot_tree é o comando responsável por gerar visualização (arquivo pdf da árvore)
tree.plot_tree(clf.fit(X, Y))
```

Figura 4 Algoritmo usado para a geração da árvore de decisão obtida pelo scikit-Learn

CONCLUSÃO

Por meio dos estudos e aplicação dos conceitos de aprendizado de máquina e mineração de dados, é perceptível que as requisições de uso dos programas do portal de Bioinformática a serem alocadas nos recursos do SDumont podem ser definidas de maneira mais eficiente, baseados nos modelos de decisão. Esses resultados reforçaram o uso dos recursos disponíveis de uma maneira mais proveitosa, trazendo mais benefícios à comunidade científica e dando um feedback aos administradores do SDumont para realizarem estudos pertinentes.

REFERÊNCIAS

Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and RegressionTrees. Wadsworth and Brooks, Monterey, CA (1984)

Demsar, J., Curk, T., Erjavec, A.,Crt Gorup, Hocevar, T., Milutinovic, M.,Mozina, M., Polajnar, M., Toplak, M., Staric, A.,Stajdohar, M., Umek, L.,Zagar, L.,Zbontar, J.,Zitnik, M., Zupan, B. Orange: Data mining tool-box in python. Journal of Machine Learning Research14, 2349–2353 (2013), http://jmlr.org/papers/v14/demsar13a.html

Hey, T., Tansley, S., Tolle, K. (eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research (2009)

Weiss, S., Kulikowski, C.: Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1991)

https://scikit-learn.org/stable/index.html

Relatório final de Iniciação Científica

Junior do Nascimento Xavier Xavier

19 de Julho de 2019

1 Dados gerais

Título do projeto: Modelagem e controle de um manipulador robótico sujeito a falhas
Bolsista: Junior do Nascimento Xavier
Orientador: DSc, Marcos Garcia Todorov
Tipo de bolsa e período: PIBIC – 08/2018 até 07/2019

2 Objetivos

- Analisar a modelagem dos manipuladores robóticos;
- Analisar a modelagem destes manipuladores no conceito das falhas;
- Analisar as ocorrências de falhas possíveis;
- Simular as falhas que podem ocorrer nos manipuladores.

3 Introdução

Manipuladores robóticos são objetos fundamentais no atual avanço do cenário tecnológico mundial, é possível encontra-los com facilidade nas indústrias do nosso dia-a-dia; e, até mesmo, no campo da medicina, entre outras diversas áreas. Tais manipuladores, visam automatizar e acelerar os processos fabris, e suas aplicações variam de acordo com a necessidade. Em áreas industriais, por exemplo, são capazes não só de agilizar os processos, como também, permitem maior padronização dos produtos e precisão, são capazes também de substituir seres humanos em tarefas de maior periculosidade. As consequências das falhas em um elemento robótico são graves, tendo em vista

que geralmente isto leva à pausa completa ou parcial das linhas de produção, gerando elevadas consequências financeiras às empresas. Neste projeto foi possível analisar primeiramente a modelagem dos manipuladores robóticos compostos por 3 links (elos), quando os mesmos não apresentam falhas, ou seja, estão em perfeito funcionamento; e quando apresentam falhas, ou seja, possuem falhas de atuação. Posteriormente, foram feitas simulações para corroborar aquilo que estava sendo estudado, permitindo uma análise gráfica das falhas e seu efeito no manipulador.

4 Materiais e Metodologia

O manipulador robótico, assim como qualquer outro elemento mecânico, está sujeito a falhas. Tais manipuladores, são compostos de forma simplificada por: elos, juntas e efetuadores. Neste projeto, a análise de falhas se dá através de falhas de atuação nas juntas do manipulador. Quando uma junta apresenta uma falha, esta junta comporta-se como junta passiva. Este efeito de passividade, permite que o controle desta junta em mau funcionamento, seja através de acoplamento dinâmico, onde é introduzido torques nas juntas ativas (juntas em perfeito funcionamento), realizando o posicionamento prévio da junta passiva até o seu *setpoint* (ponto de ajuste). Inicialmente, foi estudado a complexa modelagem do manipulador utilizando as bibliografias ([Craig, 1986], [Siqueira, 2011], [Spong, 2006], [Siqueira, 2004], [Mataric, 2017], [Spong, 2014]), quando este não apresenta falhas e, posteriormente, quando o mesmo apresenta falhas de atuação. Foi analisado as falhas que podem ocorrer nos manipuladores de 3 lins (elos) que foram os objetivos de estudo ao longo da pesquisa, e, por último, foram realizadas as análises das falhas passíveis de simulações no ambiente CERob. Para analisar graficamente o comportamento do manipulador sob o efeito de uma falha, utilizou-se um ambiente desenvolvido para o software *Matlab*, conhecido como *Control Environment of Robots* (CERob), este ambiente é disponibilizado através do livro: *Robust Control of Robots*. Neste ambiente é possível simular e, posteriormente, analisar diversas configurações de falhas nos manipuladores, pois, o mesmo dá ao usuário liberdade para definir as características físicas do manipulador; e o *setpoint* (ponto de ajuste) desejado pelo usuário. Após todo entendimento do assunto, e todas as dúvidas sanadas junto ao meu orientador, foi desenvolvido por mim junto a ele, um pôster para apresentação dos assuntos abordados nesta seção, na Jornada de Iniciação Científica e no programa de verão do LNCC.

5 Resultados e Discussão

Os resultados obtidos foram de acordo com o esperado, a pesquisa possibilitou meu aprendizado no campo da robótica, área da qual estou me graduando. Permitiu que eu pudesse conhecer melhor o que um pesquisador realiza no seu dia-a-dia de trabalho, lidar com as frustrações e prazos. Permitiu que eu obtivesse um contato com a pós-graduação do LNCC, pois fiz algumas matérias como ouvinte. A pesquisa possibilitou-me também que, o tema do qual desenvolvi no LNCC, pudesse ser ainda mais abrangido possibilitando-me levá-lo à minha universidade como tema do meu trabalho de conclusão de curso, onde pude apresentar os mesmos conceitos pouco conhecidos, que eu desenvolvi no laboratório. Por motivo de estar completando minha graduação, solicito que a bolsa não seja renovada.

6 Conclusão

O objetivo principal foi atingido, foi desenvolvido um tema de meu interesse, vivenciar o contato com pesquisadores altamente qualificados em um ambiente altamente prestigiado, entender o dia-a-dia do pesquisador, suas conquistas e frustrações, e pude me interessar ainda mais pelo trabalho de um pesquisador, além de me motivar a continuar com o desenvolvimento do tema em uma pós-graduação. Pude usar minha pesquisa como tema de trabalho de conclusão de curso na minha graduação em Engenharia Mecatrônica, o que me permitiu apresentar estes conceitos pouco conhecidos, para outras pessoas, e assim, despertar o interesse de outros na área.

Referências

[Craig, 1986] Craig, J. J. (1986). *Introduction to Robotics: mechanics and control*. Addinson-Wesley, 1 edition.

[Mataric, 2017] Mataric, M. J. (2017). *The Robotics Primer*. Massachusetts Institute of Technology, 1 edition.

[Siqueira, 2004] Siqueira, A. A. G. (2004). *Controle H_∞ não linear de robôs manipuladores subatuados*. PhD thesis, Universidade de São Paulo.

[Siqueira, 2011] Siqueira, Adriano A. G.; Terra, M. H. B. M. (2011). *Robust Control of Robots: Fault Tolerant Approaches*. Springer-Verlag London Limited, 1 edition.

[Spong, 2006] Spong, Mark W.; Hutchinson, S. V. M. (2006). *Robot Modeling and Control*. Jonh Wiley and Sons, Inc., 1 edition.

[Spong, 2014] Spong, Mark W.; Hutchinson, S. V. M. (2014). *Robot Dynamics and Control*. Jonh Wiley Sons, Inc., 2 edition.

Algoritmos Bio-Inspirados e Redes Neurais Artificiais Aplicados à Segmentação de Imagens Médicas e Biológicas

Aluno: Lucas Pampolin Laheras

Orientador: Gilson Antonio Giraldi (gilson@lncc.br)
Co-Orientador: Paulo Sérgio Silva Rodrigues (psergio@fei.edu.br)

RELATÓRIO PARCIAL

Período: 01/05/2019 a 31/07/2019
Tipo de Bolsa: Iniciação Científica

São Bernardo do Campo, SP

Julho de 2019

1 Introdução

Com o desenvolvimento de tecnologia digital, o armazenamento e gerenciamento de informações médicas e biológicas coletadas manualmente tem sido cada vez menos comuns, perdendo o lugar para as tecnologias relacionadas à digitalização automática, tanto em hardware quanto em software. Por conta disso, o desenvolvimento das áreas de processamento de imagens, inteligência artificial, aprendizagem profunda e ciência de dados é estratégico, particularmente para a área médica, uma vez que atualmente tanto hardware quanto software têm influenciado em diagnósticos, tratamentos e planejamento cirúrgico [1].

Como exemplo de aplicações onde este problema é recorrente, podemos citar a contagem de células cancerígenas, geralmente obtidas em microscópios digitais através da utilização de marcadores químicos, bem como segmentação de imagens obtidas por microscópio confocal por varredura laser, a qual vem possibilitando a análise visual de células bem como de fenômenos biológicos associados [16].

Recentemente, a literatura de processamento digital de imagens e visão computacional tem demonstrado a crescente efetividade da aplicação de algoritmos bio-inspirados para segmentação de imagem [10, 13, 18].

Por outro lado, métodos em aprendizagem profunda, baseados em rede neuronal convolucional (*Convolutional Neural Network* - CNN) foram também utilizados para segmentação de imagens, obtendo resultados muito promissores [15]. Vale ressaltar que a área de aprendizagem profunda vem ganhado espaço em inúmeras aplicações em reconhecimento de padrões, jogos, veículos autônomos, além de outras [12]. Porém, o treinamento de uma CNN para segmentação, em geral, necessita de grandes bancos e imagens contendo os dados originais e o padrão-ouro (imagem segmentada por especialista), que na maioria das vezes não estão disponíveis.

1.1 Objetivos

O objetivo do projeto é propor e analisar uma metodologia para segmentação de imagens de microscopia confocal e imagens de ultra-sonografia de mama baseada em técnicas de visão computacional utilizando algoritmos bio-inspirados para auxiliar o treinamento de uma CNN) na tarefa de segmentação de imagens.

2 Material e Métodos

2.1 Firefly

O algoritmo do *firefly*, proposto por Xin-She Yang em [19], é uma meta-heurística inspirada pelo comportamento dos vaga-lumes, que são atraídos um pelo outro de acordo com sua luminescência natural. No final, a convergência é alcançada gerando aglomerados de vaga-lumes, onde os mais brilhantes atraem os outros vaga-lumes. Se um vaga-lume em particular é o mais brilhante, ele se moverá aleatoriamente.

A ideia geral é modelar um problema de otimização não linear associando a variável de cada problema a um vaga-lume e fazer a avaliação objetiva dependendo dessas variáveis, que estão associadas a intensidade do brilho dos vaga-lumes. Então, iterativamente, as variáveis são atualizadas (seus brilhos) sob regras preestabelecidas até a convergência para um mínimo global. Genericamente, em cada geração, são realizadas as seguintes etapas: avaliar o brilho, calcular todas as distâncias entre cada par de vaga-lumes, mover cada um dos vaga-lumes em direção aos outros de acordo com o brilho, manter a melhor solução (o vaga-lume mais brilhante) e gerar aleatoriamente novas soluções.

O núcleo do algoritmo é sua função de avaliação Z, que depende do problema atual, especificamente o problema de limiarização de vários níveis (*MultiLevel Thresholding Problem* - MLOTP), cada vaga-lume é considerado uma variável d-dimensional, onde para cada dimensão é computado um limiar distinto, particionando o espaço do histograma.

2.2 Redes Neuronais Convolucionais (CNN)

As CNNs têm se mostrado atrativas desde o ano de 2012 [7], em tarefas complexas de classificação de imagens. Inicialmente propostas por LeCun *et al.* [8], trata-se da aplicação de aprendizado profundo, em que redes neurais de muitas camadas são aplicadas em imagens. Porém, diferentemente de redes neurais comuns, a principal operação que ocorre nas imagens é a chamada convolução. Uma imagem gerada pela convolução de um filtro f com outra imagem I é definida por $g(m,n) = \sum_{i=1}^{j} \sum_{j=1}^{q} f(i,j)\, I(m-i, n-j)$.

As CNNs são compostas por cinco camadas básicas. Camada de entrada, onde a imagem é inserida na rede. Camada de Convolução, que contém diversos filtros lineares de tamanho fixo usados para realizarem a convolução, gerando o chamado mapa característico da imagem que ressaltam padrões na imagem. Camada RELU, ocorre após a operação de convolução com aplicação de uma função não linear à saída x da camada anterior. De acordo com Krizhevsky *et al.* [7], elas são usadas para a rede convergir mais rápido. Camada de Subamostragem, a qual sumariza os dados ao deslizar uma janela nos mapas

de características gerados pelas camadas convolucionais, aplicando operações para criar um novo mapa de características. Camadas totalmente conectadas, localizadas no fim da rede agindo como classificadores, geralmente são do tipo soft-max usados para determinar a classe associada à imagem de entrada.

Os filtros das camadas convolucionais das CNNs são definidos por máscaras com tamanhos reduzidos, permitindo uma extração eficiente de características de alto nível que são utilizadas pelas camadas totalmente conectadas. O treinamento de uma CNN é feito por propagação reversa do erro via descida do gradiente estocástico. Erros de classificação na etapa de treinamento são usados para atualizar os pesos das máscaras de convolução das camadas convolucionais e das camadas totalmente conectadas.

Além da segmentação, as redes CNN vêm sendo aplicadas para extração de características em imagens de dígitos [5], textura [3], detecção de objetos em imagens e vídeos [4], dentre outras aplicações [2, 6].

2.3 Algoritmo Level Sets

O conceito básico do algoritmo é delimitar a fronteira de uma região planar utilizando o conjunto de pontos no nível zero de uma função $\alpha : \Re^2 \times \Re^+ \rightarrow \Re$ dado pela seguinte expressão:

$$S(t) = \left\{(x, y) \in \Re^2 \mid \alpha(x, y, t) = 0\right\} \tag{1}$$

Dada uma curva de nível zero $S(t)$ no instante t, consideram-se três possíveis valores para a função. Quando o valor é menor que 0 a coordenada (x, y) está dentro da superfície S, quando o valor é maior que 0 a coordenada (x, y) está fora da superfície e quando o valor for igual a 0 a coordenada (x, y) é o contorno da superfície.

O próximo passo é encontrar a formulação euleriana da evolução da frente. gerando uma equação diferencial parcial que descreve a evolução da curva de nível no tempo e no espaço [9].

2.4 Banco de Dados

Os bancos de imagens biológicas e médicas para os testes dos algoritmos serão formados por imagens de microscopia confocal e por imagens de ultra-sonografia de mama, respectivamente. No primeiro caso, vamos aproveitar o banco de imagens utilizado em um trabalho de iniciação científica concluído no ano passado, o qual recebeu destaque em Jornada de iniciação científica, ano 2018 [11]. Com relação as imagens médicas, utilizaremos

o banco de imagens de ultra-sonografia disponível na equipe do Centro Universitário da FEI, o qual foi utilizado em artigo publicado recentemente [14]. As implementações estão sendo realizadas na linguagem Python. Em relação às técnicas de CNN, utilizaremos a biblioteca de código livre denominada TensorFlow [17], também disponíveis em Python. Daremos continuidade ao desenvolvimento de software, tanto em relação a CNN quanto aos algorítimos bio-inspirados.

2.5 Metodologia

O presente projeto envolve as seguintes etapas: (a) Desenvolvimento de uma plataforma interativa para segmentação das imagens de interesse; (b) Treinamento tradicional de CNN; (c) Utilização do software desenvolvido na primeira etapa para melhorar o treinamento da CNN.

A etapa (a) está fundamentada em algoritmos (Seção 2.1), podendo também utilizar técnicas baseadas em métodos de contorno ativo, tais como level sets (Seção 2.3). Neste passo, serão aplicadas técnicas já implementadas pelas equipes envolvidas, as quais serão integradas em um software com interfaces e recursos gráficos específicos para este projeto. Na etapa (b) será utilizado um modelo de CNN desenvolvido recentemente pelas equipes do LNCC e da FEI, utilizado para gerar alguns resultados publicados em [14]. O passo (c) dependera da integração do software implementado na primeira etapa com o treinamento da CNN.

Vale ressaltar que a área de aprendizagem profunda baseada em modelos to tipo CNN é o estado-da-arte em termos de reconhecimento de padrões e que a metodologia proposta poderá ser utilizada para outras aplicações fora do campo de segmentação de imagens, o que é mais uma motivação para o presente projeto.

3 Resultados Parciais

O pipeline exibido na Figura 1 foi implementado em Python, sendo composto por 6 módulos principais, descritos em mais detalhes a seguir.

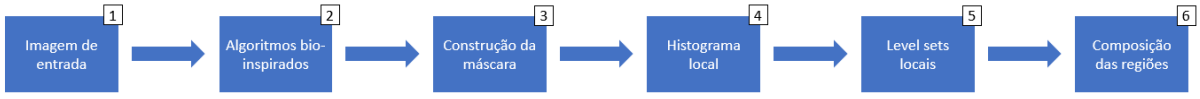


Fig. 1: Pipeline utilizado em estudos iniciais.

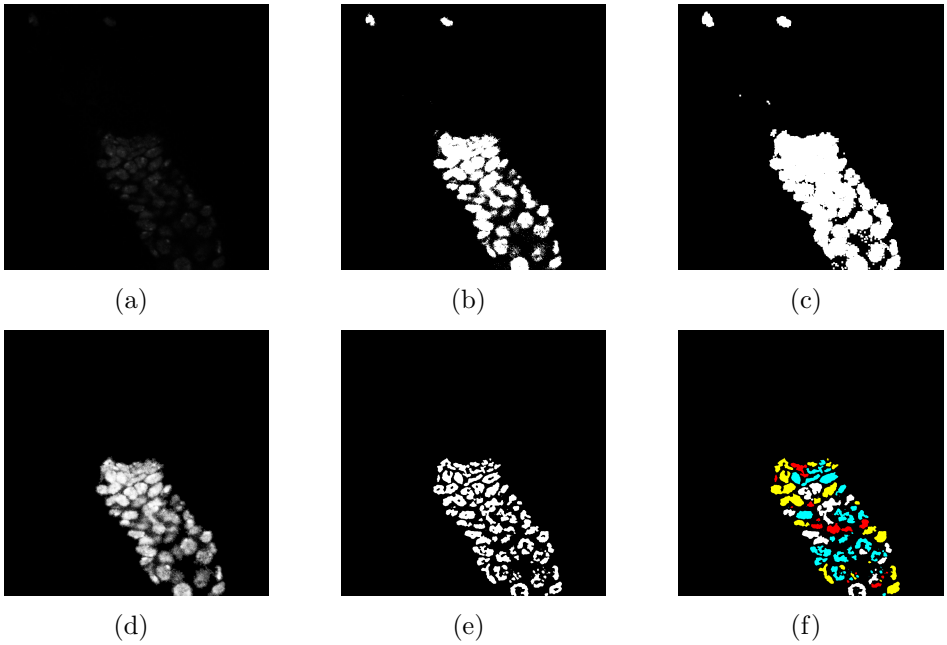


Fig. 2: (a) Imagem original. (b) Firefly aplicado na imagem após ajuste de contraste. (c) Mascaras delimitando as regiões obtidas. (d) Equalização do histograma aplicado localmente. (e) Resultado da aplicação do método de Level set 2.3. (f) Composição e coloração das regiões.

O primeiro bloco (bloco 1, Figura 1) altera a imagem original, mostrada na Figura 2.(a), para tons-de-cinza. Pode-se observar o fraco contraste entre os objetos e o fundo da imagem, o que dificulta muito o tratamento automático destas imagens.

Na etapa correspondente ao bloco 2 da Figura 1, a imagem fornecida pela etapa 1 é processada pelo algoritmo bio-inspirado *Firefly* (Seção 2.1). A saída dessa imagem é mostrada na Figura 2.(b).

No passo correspondente à construção da máscara (bloco 3, Figura 1) a imagem de saída da etapa 2 é composta por $N \geq 1$ regiões de interesse. Se a área de uma dessas regiões for menor que um limiar $L > 0, L \in \mathbb{N}$, então essa região é definida como sendo uma máscara binária. Sendo assim, a saída dessa etapa é o conjunto $S = \{m_1, m_2, ..., m_N\}$ de máscaras binárias. Esta etapa esta representada na Figura 2.(c).

Cada máscara binária definida na etapa anterior é utilizada para a etapa de histograma local (Bloco 4, Figura 1), sobre a imagem original para delimitar a região onde uma equalização local será executada. A razão da aplicação da equalização local é o fato de que, localmente, as regiões inomogêneas são reduzidas. A Figura 2.(d) é o resultado equalização da Figura 2.(a) utilizando a região branca da Figura 2.(c).

Na etapa referente ao bloco 5 da Figura 1, para cada imagem de saída da etapa

anterior é utilizado o algoritmo level sets (Seção 2.3) de forma a realçar e delimitar as regiões locais, para obter um resultado melhor na contagem de células, como pode ser visto na Figura 2.(e).

Na última etapa (bloco 6, Figura 1), será feito o agrupamento de cada imagem de saída da etapa anterior em uma única imagem. Após o agrupamento, a cada região é atribuída uma cor aleatoriamente, de maneira que possam ser distinguidas visualmente, como pode ser observado na Figura 2.(f).

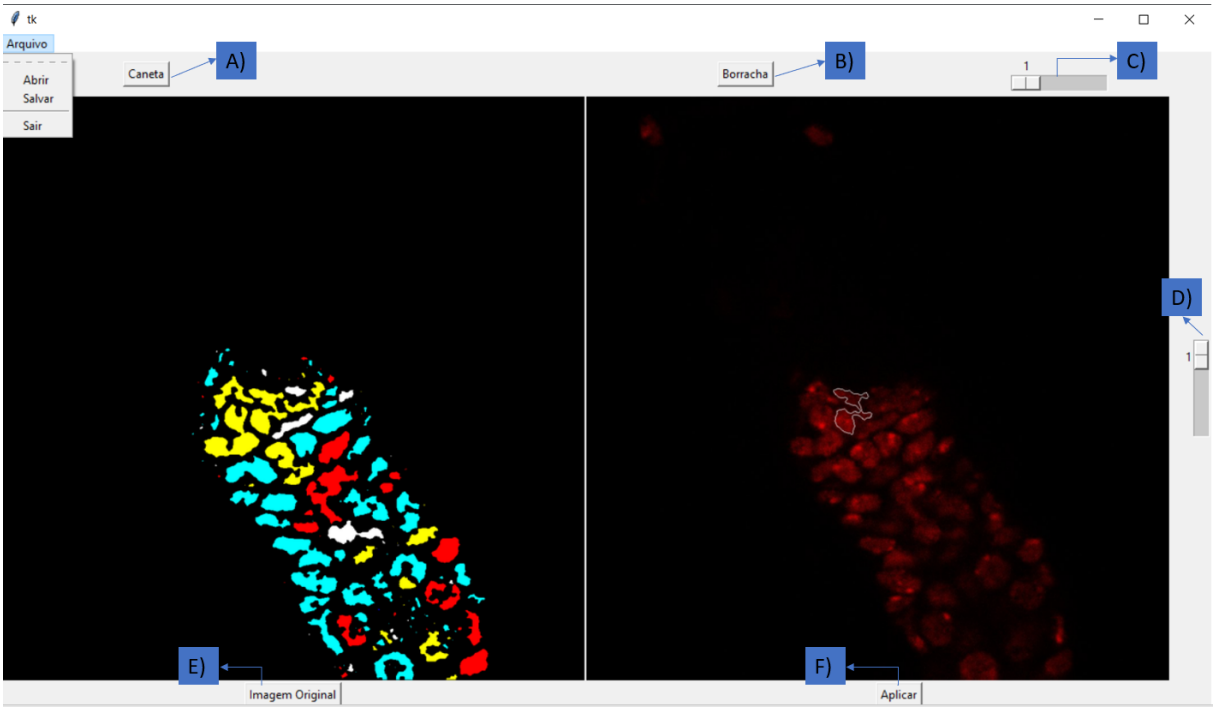


Fig. 3: Layout da interface gráfica desenvolvida.

O *pipeline* de segmentação descrito acima foi implementado em uma plataforma interativa (Figura 3) para que os biólogos e médicos das equipes envolvidas possam determinar o padrão ouro das imagens que farão parte do banco de dados utilizado para o treinamento da CNN.

Para utilizar essas funções é preciso clicar na imagem da esquerda na região que está segmentada de forma errada e com *mouse* passar por cima da área da imagem à direita com a opção caneta ou borracha selecionada (A e B da Figura 3) e após concluído clicar em "Aplicar"(botão referente a letra F na Figura 3), assim, modificando a imagem da esquerda para obter a segmentação correta. A barra de rolagem C determina o tamanho da borracha ou caneta selecionada. O botão E altera a imagem da esquerda para a imagem que foi inicialmente gerada pelo algorítimo. Para facilitar para o usuário demarcar a

região, está sendo implementada a função de zoom da imagem utilizando a barra de rolagem D.

4 Conclusão

Este relatório apresenta os resultados parciais do projeto definido na Seção 3. Foi implementada uma interface que está na sua versão beta (Figura 3) e encontra-se em teste. Esta interface executa o algoritmo de segmentação e com base na imagem gerada, o usuário seleciona regiões que precisam ser corrigidas.

Para os próximos passos será feito o seguinte cronograma:

- Implementação de redes neurais dentro do TensorFlow: 01/07/2019 a 30/09/2019
- Segmentação de imagens via CNN: 01/09/2019 a 30/11/2019
- Implementação de software para treinamento de CNN com supervisão do usuário: 01/12/2019 a 30/02/2020
- Geração de resultados e Submissão de artigo científico: 01/03/2020 a 30/04/2020

Referências

[1] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211, 2007.

[2] PN Druzhkov and VD Kustikova. A survey of deep learning methods and software tools for image classification and object detection. *Pattern Recognition and Image Analysis*, 26(1):9–15, 2016.

[3] Anselmo Ferreira and Gilson Giraldi. Convolutional neural network approaches to granite tiles classification. *Expert Systems with Applications*, 84:1–11, 2017.

[4] Christophe Garcia and Manolis Delakis. Convolutional face finder: A neural architecture for fast and robust face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 26(11):1408–1423, 2004.

[5] Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*, 2013.

[6] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[9] Ravi Malladi, James A Sethian, and Baba C Vemuri. Shape modeling with front propagation: A level set approach. 1994.

[10] Hongwei Mo, Lifang Xu, and Mengjiao Geng. Image segmentation based on bio-inspired optimization algorithms. pages 259–284, 01 2014.

[11] M. A. D. MOURA and G. A. Giraldi. Destaque na xiii jornada de iniciaÃ§Ã£o cientÃfica e tecnolÃ³gica do lncc. 2018.

[12] Philosophia Naturalis. 30 amazing applications of deep learning, Jun 2018.

[13] Valentín Osuna-Enciso. Bioinspired metaheuristics for image segmentation. 2014.

[14] Paulo Sergio Rodrigues, Guilherme Wachs-Lopes, Ricardo Morello Santos, Eduardo Coltri, and Gilson Antonio Giraldi. A q-extension of sigmoid functions and the application for enhancement of ultrasound images. *Entropy*, 21(4):430, 2019.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[16] David J Stephens and Victoria J Allan. Light microscopy techniques for live cell imaging. *science*, 300(5616):82–86, 2003.

[17] TensorFlow. https://www.tensorflow.org/.

[18] G. A. Wachs Lopes, F. S. Beltrame, R. M. Santos, and P. S. Rodrigues. Comparison of bio-inspired algorithms from the point of view of medical image segmentation. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–7, July 2018.

[19] Xin-She Yang. Firefly algorithms for multimodal optimization. In *SAGA*, 2009.

Laboratório Nacional de Computação Científica
Coordenação de Ciência da Computação (CCC)

Relatório de atividades

Programa Institucional de Bolsas de Iniciação Científica
PIBIC

Uso da Computação Distribuída de Alto Desempenho para Fluidodinâmica Computacional: Um Estudo de Eficiência Energética e Desempenho

Matheus de Oliveira Pires

Orientador - Prof. Bruno Schulze

Período Agosto/2018 a Maio/2019

Julho 2019

Sumário

1 Introdução **1**

2 Objetivo **1**

3 Metodologia **2**

3.1 Experimento 2

4 Resultados e Discussão **3**

5 Conclusões **4**

6 Agradecimentos **4**

Dados Referentes à Bolsa

- **Instituição:** Laboratório Nacional de Computação Científica
- **Bolsista:** Matheus de Oliveira Pires
- **Coordenador/Orientador:** Bruno Schulze (CCC-LNCC)
- **Coorientadora:** Mariza Ferro (LNCC)
- **Modalidade:** Bolsa de Iniciação Científica
- **Período de Vigência da Bolsa:** Agosto de 2018 até Maio de 2019
- **Projetos Relacionados:** ComCiDis (Computação Científica Distribuída)

1 Introdução

A computação de Alto Desempenho (HPC) vem exercendo um papel fundamental no avanço de muitas áreas das ciências, pois ela oferta uma infraestrutura computacional capaz de processar um grande volume de dados em um curto período de tempo. Porém, apesar do seu avanço nas ultimas décadas, com a chegada dos supercomputadores petaflóficos[1], ainda existem diversos domínios de aplicações que necessitam de um maior poder computacional. Entre eles está a Dinâmicas de fluidos Computacionais (CFD), que consiste na "área da computação científica que estuda métodos computacionais para simulação de fenômenos que envolvem fluidos em movimento, com ou sem troca de calor" [de Oliveira Fortuna 2000]. Esta ciência apresenta equações complexas que necessitam de técnicas computacionais que demandam um alto poder computacional. Por isso, muitas simulações ainda não alcançam a precisão desejada ou até mesmo são inviáveis de serem realizadas. Em vista disso, é necessário que exista um aumento significativo no poder computacional até atingir a exaescala[2].

Porém, uma das principais barreiras para a chegada a exaescala é a redução no consumo de energia. Em um estudo realizado pelo Departamento de Energia dos EUA, estabeleceu-se um limite de 20 MW (Megawatts) de consumo para tornar viável a futura exaescala [Kogge et al. 2008]. Diante das tecnologias atuais, isso representa um acréscimo significativo de desempenho, mas sem um aumento no consumo de energia. O que faz necessário, entender o recursos computacionais e como eles estão relacionados com o consumo de energia. Assim será possível aproveitar melhor os recurso disponíveis e identificação quais são os principais fatores que estão limitando o desempenho da aplicação.

2 Objetivo

O objetivo desse trabalho é estudar o desempenho computacional e o consumo de energia de aplicações cientificas na área de dinâmicas dos fluidos computacional e compreender como o modelo matemático é implementado, utilizando uma aplicação real (uma simulação CFD). Além disso, propor um modelo de predição dos recursos computacionais e energia da aplicação estudada. Assim, espera-se contribuir para a definição de otimizações que permitam alcançar bom desempenho e escalabilidade dessas aplicações cientificas, bem como eficiência no consumo energético.

[1]Petaflópicos - Sistemas que podem executar 10^{15} operações de ponto flutuante por segundo.
[2]ExaFlop - 10^{18} operações de ponto flutuante por segundo.

3 Metodologia

Para cumprir os objetivos acima descritos foi adotada uma metodologia para o desenvolvimento da pesquisa.

- Estudo da ferramenta *perf*, com a qual foi possível obter uma série de parâmetros importantes, tais como informações de CPU (cpu-cycles, cpu-clock ,etc), memória (cache misses, cache-references) e energia.
- Revisão bibliográfica ([Hennessy and Patterson 2014] que possibilitou compreender os dados coletados pelo perf.
- Estudo do software de simulação de fluidodinâmica OpenFOAM[3], o qual possibilita resolver equações diferenciais parciais pelo método dos volumes finitos. Foi estudado o modelo de volumes finitos e a estrutura do programa (que é implementado em C++.) para que fosse possível utilizá-lo.
- Estudo da literatura [de Andrade Silva 2017] para compreender alguns conceitos de hemodinâmica. Devido à complexidade, ainda não foi possível realizar uma simulação de escoamento sanguíneo conforme previsto no plano de trabalho. Além disso ainda não foi encontrada uma aplicação real implementada no openFOAM que esteja relacionada ao escoamento sanguíneo.
- Estudo dos modelos do escoamento em uma cavidade e tutorial relacionado a paralelização com MPI.
- Estudo da ferramenta R utilizada para analisar os dados e realizar a regressão dos dados coletados.

3.1 Experimento

Com objetivo de entender os recurso computacionais, foi executado uma simulação de escoamento em uma cavidade utilizando o software openFOAM, de código aberto e paralelizado com MPI conforme especificado no tutorial [4].

As hipóteses utilizadas para simplificar o problema foram:

- Escoamento transiente
- Escoamento incompressível
- Escoamento no plano
- Escoamento laminar
- Isotérmico

O modelo utilizado é a equação de Navier-Stokes para fluidos incompressíveis[5].

$$div(\vec{u}) = 0 \tag{1}$$

[3] https://www.openfoam.com/
[4] https://cfd.direct/openfoam/user-guide/v6-running-applications-parallel/
[5] https://www.openfoam.com/documentation/guides/latest/api/icoFoam_8C.html

$$\frac{\partial}{\partial t}(U) + \nabla \cdot (UU) - \nabla \cdot (\nu \nabla U) = -\nabla p \tag{2}$$

No experimento, foram executados 30 vezes cada tamanho de malha, variando de 5 subdivisões até 100, paralelizado com 1, 2, 4 e 8 processos, com número de número de Reynolds 10 e 1000. Foram coletados 27 parâmetros de desempenho computacional e energia através da ferramenta de monitoramento perf. Além disso, foi necessário mudar a propriedade da viscosidade para aumentar o número de Reynolds para 1000, calcular o a variação do tempo para manter o número de Courant igual a 1, conforme estabelecido pelo tutorial [6].

Realizou-se o cálculo da média, do desvio padrão de cada tamanho de malha e utilizou-se um modelo de regressão polinomial para ajustar um modelo de predição dos recursos computacionais e energia pelo tamanho de subdivisões da malha.

4 Resultados e Discussão

Na Figura 1 é apresentada a evolução dos recursos computacionais coletados à medida que vai aumentando a subdivisão de malha e aplicando o paralelismo com MPI para a configuração do número de Reynolds igual 10. A Figura 2 contém os resultados para o número de Reynolds 1000. Nas Figuras 1 e 2 os gráficos apresentados nas colunas da esquerda mostram os eventos coletados para um processo MPI, enquanto os da coluna da direita os mesmos experimentos, porém, paralelizados em 8 processos.

Pode-se perceber que o consumo dos recursos computacionais nos experimentos apresentados na Figura 2 são maiores do que os apresentados na Figura 1. Isso se deve ao aumento da vorticidade do escoamento.

Os eventos *instructions* (1a, 1b, 2a, 2b), *cache misses* (1c, 1d, 2c, 2d) e *energy cores* (1g, 1h, 2g, 2h) tiveram um acréscimo significativo com aumento das subdivisões da malha e o número de processos. O tempo de execução da aplicação aumentou com o aumento da malha. Porém, o tempo de execução diminui até o quarto processo na paralelização com MPI, mas a partir do oitavo processo não houve mais ganhos e o tempo de execução aumentou.

Em relação aos eventos para análise de consumo de energia, foram analisados o consumo dos núcleos do processador (*energy cores*), da memória RAM (energy ram) e GPU (energy gpu). O evento que apresentou o maior consumo de energia entre os eventos coletados é o *energy cores*. Isso se deve ao aumento no número de *instructions*, *cpu-clock* e *cycles*. Ou seja, os principais eventos relacionados ao uso da CPU, devido ao aumento do número de processos e como consequência maior utilização dos núcleos de processamento.

Para cada um dos gráficos pode ser observado que os modelos de regressão polinomial se ajustaram bem aos valores coletados. Porém, alguns eventos apresentaram um alto desvio padrão e outros não se adequaram ao modelo de regressão utilizado (por conta de espaço não foi possível colocar, mas estão disponíveis todos os 27 parâmetros em:[7]). O erro sobre os modelos de regressão ainda não foram avaliados, pois são necessárias novas execuções de teste, com tamanhos de malha maiores para verificar esses valores.

[6] https://www.openfoam.com/documentation/tutorial-guide/tutorialse2.phpx6-60002.1
[7] https://github.com/matheus812/experimento$_o penfoam$

5 Conclusões

Com a finalidade de contribuir com pesquisas em computação científica que buscam soluções para viabilizar uma nova geração de supercomputadores, é necessário compreender os recursos computacionais e como eles estão relacionados com o consumo de energia. Assim, é possível encontrar os gargalos nas execuções. Para isso foi estudado o desempenho computacional e o consumo de energia de aplicações cientificas na área de dinâmicas dos fluidos computacional e foi feita uma simulação de escoamento em uma cavidade e coletados os eventos relacionados ao consumo de recursos computacionais. Com os resultados dos experimentos foram desenvolvidos modelos de regressão polinomial com o objetivo de predizer o comportamento para problemas mais complexos e compreender os requisitos computacionais desses modelos CFD. Analisando os resultados é possível observar que conforme se aumenta o número de Reynolds, representando escoamentos menos laminares, a complexidade para resolver o problema também aumenta e consequentemente o consumo de recursos, especialmente de CPU, aumentando o consumo de energia. Apesar do resultado ser bastante óbvio, a metodologia se mostra eficaz para estudar como a relação do aumento de complexidade do problema (Reynolds e a subdivisão da malha) e a análise da taxa de crescimento das curvas de consumo de recursos pode ser usada para resultados práticos em trabalhos futuros. Os modelos de regressão polinomial se ajustaram bem aos valores coletados. Porém, alguns eventos apresentaram um alto desvio padrão e outros não se adequaram ao modelo de regressão utilizado. O erro sobre os modelos de regressão ainda não foram avaliados, pois são necessárias novas execuções de teste, com tamanhos de malha maiores para verificar esses valores. Além disso, para um melhor ajuste das curvas é interessante avaliar outros modelos de regressão.

6 Agradecimentos

Agradeço ao Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq) pelo apoio financeiro e ao Laboratório Nacional de Computação Científica (LNCC) pela oportunidade concedida.

Referências

[de Andrade Silva 2017] de Andrade Silva, J. (2017). *Modelagem computacional de escoamento sanguıneo em fıstulas arteriovenosas para o processo de hemodiálise*. PhD thesis, Universidade Federal de Juiz de Fora. 2

[de Oliveira Fortuna 2000] de Oliveira Fortuna, A. (2000). *Técnicas computacionais para dinâminca dos fluidos: conceitos básicos e aplicações*. Edusp. 1

[Hennessy and Patterson 2014] Hennessy, J. L. and Patterson, D. A. (2014). *Arquitetura de Computadores: Uma Abordagem Quantitativa*. Elsevier Editora Ltda, fifth edition. 2

[Kogge et al. 2008] Kogge, P., Bergman, K., Borkar, S., Campbell, D., Carlson, W., Dally, W., Denneau, M., Franzon, P., Harrod, W., Hill, K., Hiller, J., Karp, S., Keckler, S., Klein, D., Lucas, R., Richards, M., Scarpelli, A., Scott, S., Snavely, A., Sterling, T., Williams, R. S., and Yelick, K. (2008). ExaScale Computing Study: Technology Challenges in Achieving ExaScale Systems. Technical report, DARPA IPTO, Air Force Research Labs. 1

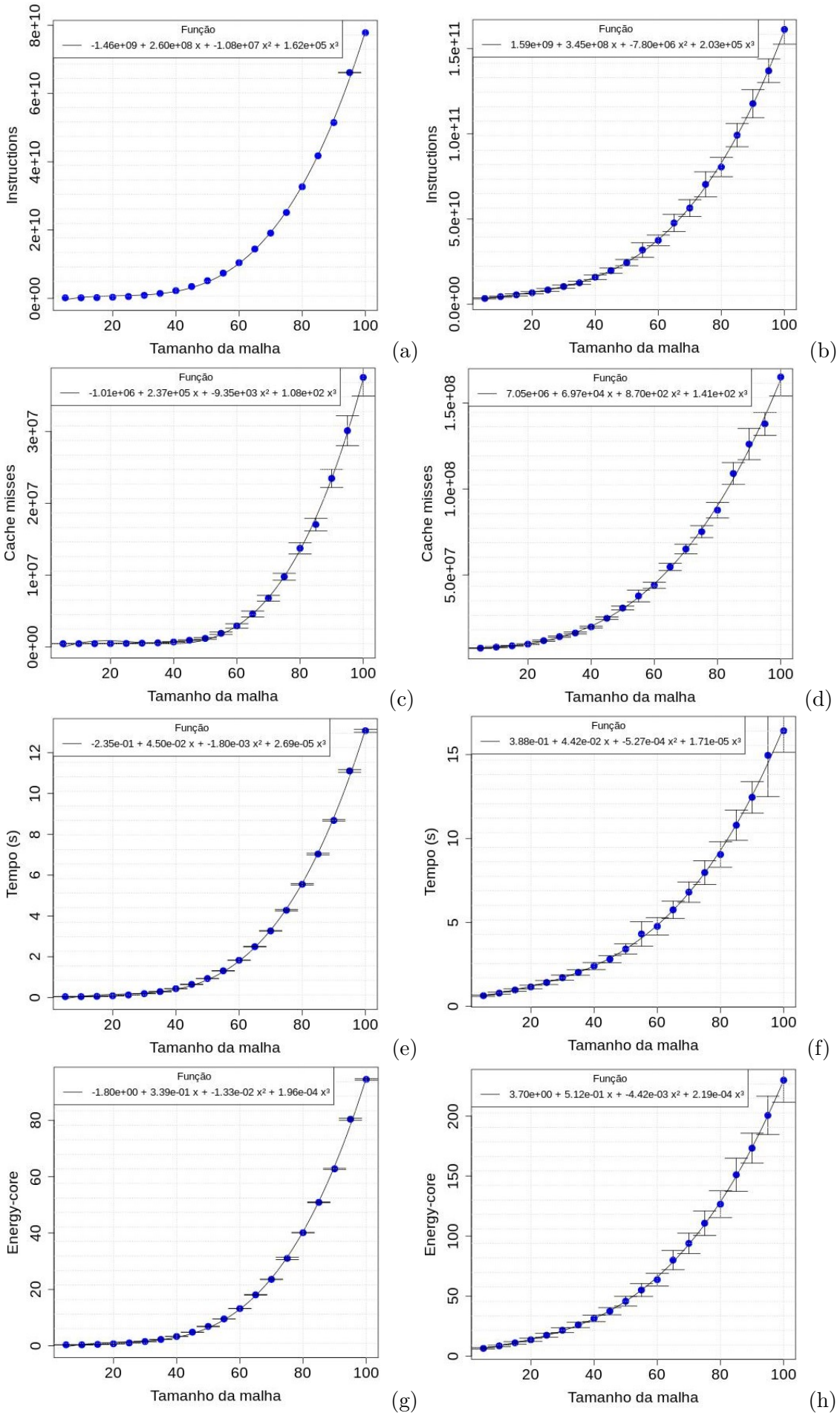


Figura 1: Eventos coletados para número de reynolds 10.

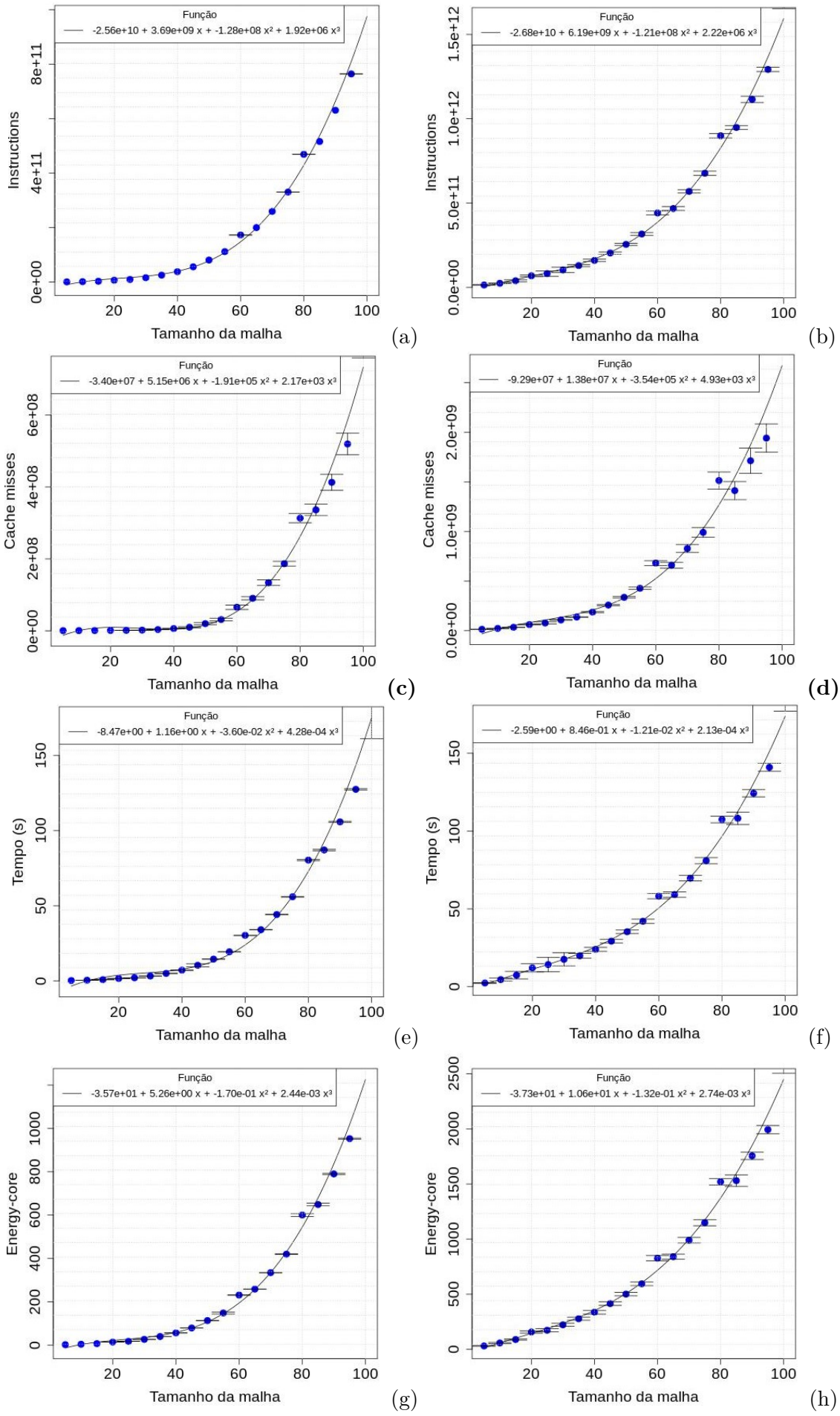


Figura 2: Eventos coletados para número de reynolds 1000.

Laboratório Nacional de Computação Científica
Coordenação de Ciência da Computação (CCC)

Relatório de atividades

Programa Institucional de Bolsas de Iniciação Científica PIBIC

Desenvolvimento de Estratégias Autonômicas para a Eficiência Energética em Ambientes HPC

Matheus Gritz Alves de Souza

Orientador - Prof. Bruno Schulze

Período Agosto/2018 a Julho/2019

Julho 2019

Sumário

1 Introdução **1**

2 Objetivo **2**

3 Metodologia **2**

4 Resultados e Discussão **3**

5 Publicações **5**

6 Conclusões **5**

7 Agradecimentos **6**

Dados Referentes à Bolsa

- **Instituição:** Laboratório Nacional de Computação Científica
- **Bolsista:** Matheus Gritz Alves de Souza
- **Coordenador/Orientador:** Bruno Schulze (CCC-LNCC)
- **Coorientadora:** Mariza Ferro (LNCC)
- **Modalidade:** Bolsa de Iniciação Científica
- **Período de Vigência da Bolsa:** Agosto de 2018 até Julho de 2019
- **Projetos Relacionados:** ComCiDis (Computação Científica Distribuída)

1 Introdução

A Computação de Alto Desempenho (HPC) tem se tornado crucial para as pesquisas científicas em muitos domínios de investigação. Porém, apesar da capacidade de processamento dos atuais supercomputadores petaflóficos, diversos domínios de aplicação ainda necessitam de maior poder computacional, como o esperado para a futura exaescala de processamento. Entre os principais desafios a serem superados para se alcançar esses níveis de processamento estão [Messina 2017], [HPC4e 2017] a preparação das próprias simulações e algoritmos para que de fato possam alcançar os níveis de processamento que serão oferecidos por essas futuras arquiteturas e a eficiência energética. Em um estudo realizado pelo Departamento de Energia dos EUA, estabeleceu-se um limite de 20 MW (Megawatts) de consumo para tornar viável a futura exaescala [Kogge et al. 2008]. Diante das tecnologias atuais, isso representa alcançar um aumento significativo de desempenho, mas sem elevar o consumo de energia.

Ainda, estes estudos apontam para possíveis estratégias que permitam superar esses desafios e, ao contrário dos sistemas passados, a próxima geração de supercomputadores precisará ser desenvolvida usando abordagens onde os requisitos do problema científico orientem a arquitetura do computador e o projeto do software do sistema. Além disso, esses requisitos do problema científico deverão orientar a orquestração de diferentes técnicas e mecanismos de economia de energia, com a finalidade de melhorar o balanceamento entre economia de energia e desempenho das aplicações. Para isso o uso de técnicas autonômicas que permitam desde o melhor escalonamento das aplicações até o dimensionamento de frequência de processadores e memórias, cientes dos requisitos das aplicações serão fundamentais.

Porém, para se alcançar esse nível de orquestração é fundamental compreender aplicações e ambientes de HPC e como seus componentes se relacionam com o consumo de energia. É preciso aprofundar o conhecimento sobre os fatores que limitam o desempenho das aplicações e interferem no consumo de energia, e mapeá-los para as arquiteturas que representam o atual estado da arte no processamento de alto desempenho.

Assim, o primeiro passo para o desenvolvimento de estratégias autonômicas, cientes das aplicações (*application aware*), é a compreensão sobre os requisitos das aplicações científicas, e o segundo, a definição dos parâmetros mais relevantes a serem monitorados sobre a execução da aplicação permitindo identificar níveis de desempenho e consumo de energia.

As aplicações científicas possuem diferentes requisitos computacionais, pois de acordo com o domínio do problema sendo modelado, diferentes métodos matemáticos são utilizados. Na abordagem proposta neste projeto, busca-se tratar esse tipo de problema avaliando o desempenho de aplicações com base em uma classificação denominada de Motifs [Patterson 2013],

a qual considera as características das aplicações científicas em termos dos seus requisitos computacionais. Assim, esta categorização é utilizada como referencial para o estudo do comportamento das aplicações científicas, identificação do núcleo básico da aplicação (onde o processamento é realmente feito) e de como elas se relacionam com outros aspectos que modifiquem o seu desempenho e consumo de energia, auxiliando na descoberta dos aspectos relevantes para se alcançar a melhoria geral do desempenho e do consumo energético.

O segundo passo é o monitoramento das aplicações e a identificação dos parâmetros mais relevantes que permitam identificar este núcleo de processamento, e sua classe de aplicação, e a orquestração do melhor balanço entre eficiência energética e desempenho. Entretanto, para permitir uma avaliação de todos esses aspectos e seus relacionamentos, outro desafio é definir uma maneira precisa de coletar todos esses parâmetros e relacioná-los para uma análise eficaz. Porém, esta não é uma tarefa trivial, pois é necessário coletar um conjunto de parâmetros distintos, que podem mudar a cada avaliação, em diferentes ambientes HPC. Para isso é proposto o estudo de como coletar os valores dos contadores de hardware que atendam aos objetivos e a identificação dos parâmetros mais relevantes para o problema proposto neste trabalho.

2 Objetivo

Motivados pela relevância do tema e dos aspectos desafiadores a serem explorados, o objetivo geral deste projeto de pesquisa é o desenvolvimento de um sistema de gestão autonômico cientes das aplicações que permita a orquestração de diferentes técnicas e mecanismos de economia de energia, com a finalidade de melhorar o balanceamento entre economia de energia e desempenho das aplicações.

De forma mais pontual, o objetivo deste trabalho é o estudo do desempenho e consumo de energia de aplicações científicas, identificando parâmetros obtidos por meio dos contadores de hardware que permitam identificar o núcleo básico de uma aplicação, avaliar desempenho e consumo de energia. Entre os objetivos técnicos deste projeto estão:

a. Definir os parâmetros mais relevantes que permitam identificar o núcleo principal de execução das aplicações e avaliar desempenho e consumo de energia para a execução dessas aplicações;

b. Investigar e caracterizar os requisitos computacionais de aplicações/simulações científicas usando como referencial as classes de aplicações Motifs;

c. Compreender como essas classes e seus requisitos se relacionam com o consumo de energia, quais os principais fatores que limitam o desempenho dessas classes.

3 Metodologia

Para cumprir os objetivos acima descritos foi adotada uma metodologia para o desenvolvimento da pesquisa.

- Análise técnica de um conjunto de ferramentas de monitoramento de desempenho e energia, verificando suas vantagens, limitações e como essas ferramentas apresentam os dados para análise, avaliando a facilidade de instalação, utilização e interpretação destes resultados.

- Estudo da ferramenta *perf*, com a qual foi possível obter uma série de parâmetros importantes, tais como informações de CPU (instructions, cpu-cycles, etc), memória (mem-load, mem-store), cache (L1-dcache-loads, LLC-stores, etc) e energia.

- Estudo da linguagem de programação *Python 3.6*[1], incluindo seus conceitos, princípios, convenções e estruturas específicas de dados.
- Estudo da biblioteca de análise de dados *pandas*[2] que oferece uma série de estruturas de dados adicionais para a linguagem Python.
- Estudo e revisão bibliográfica da biblioteca de aprendizado de máquina *scikit-learn*[3] [Pedregosa et al. 2011] e seus diversos modelos de regressão.
- Coleta de parâmetros de desempenho com a ferramenta *perf* e análise da correlação e covariância entre eles.
- Limpeza, preprocessamento dos dados e seleção dos atributos mais relevantes para o alvo do treinamento do modelo de aprendizado de máquina selecionado.
- Desenvolvimento de um modelo de regressão baseado em Árvore de Decisão Regressora para a predição do tempo de execução e consumo energético usando os dados previamente coletados como a base do treinamento.

4 Resultados e Discussão

Seguindo as etapas descritas na Seção 3, foi feita uma comparação entre as principais ferramentas de coleta de desempenho disponíveis. Na Tabela 1 é apresentado um comparativo entre elas.

App	Desempenho			Energia		
	CPU	Memória	I/O	PKG	Cores	RAM
perf	Sim	Não	Parcial	Sim	Sim	Sim
likwid	Sim	Sim	Não	Sim	Sim	Sim
hpctoolkit	Sim	Sim	Sim	Não	Não	Não

Tabela 1: Comparativo de Ferramentas

A ferramenta selecionada por sua vasta quantidade e variedade de parâmetros disponíveis e que atende a maior parte dos requisitos é o *perf*, disponível no kernel do Linux por meio do pacote *linux-tools-generic*. O perf é uma ferramenta de coleta de contadores de hardware que utiliza o registradores específicos da arquitetura (*MSR*) e retorna um total baseado na contagem total ou uma amostra baseado em intervalo de tempo. Na Tabela 2 é detalhado os contadores que foram coletados pelo perf e utilizados como atributos para tarefa de aprendizado neste trabalho.

Após uma extensiva análise, foi possível identificar relações entre os diversos contadores de desempenho a fim de elaborar o desenvolvimento de um modelo de aprendizado de máquina para a predição de parâmetros referentes à tempo de execução e consumo energético.

O mapa de calor na Figura 1 mostra a correlação entre os atributos coletados. É possível observar que parâmetros associados ao processamento (*instructions*, *branches*) e ao cache (*L1-dcache-stores*, *L1-dcache-loads*, *L1-dcache-load-misses*) tem um impacto maior, tanto no tempo de execução (*runtime*), quanto no consumo energético (*energy_joules*).

O ambiente de coleta dos dados foi feito inteiramente em um dos nós do cluster do grupo ComCiDis. Ele abriga dois processadores Intel Xeon X5650S de seis núcleos cada com o *clock*

[1]https://www.python.org/
[2]https://pandas.pydata.org/
[3]https://scikit-learn.org/

	Contadores/Atributos	Descrição
1	Instructions	# de instruções enviadas a CPU
2	Cycles	# de ciclos de CPU completados
3	CPU Migrations	# de vezes que o processo trocou de CPU
4	Branches	# de instruções condicionais que alteram o fluxo da execução
5	Branch Misses	# de vezes que a CPU falhou em predizer uma instrução condicional
6	Context Switches	# de vezes que o processo parou para que outro executasse
7	Cache References	# de vezes que um dos níveis de cache foi acessado
8	Cache Misses	# de vezes que algo não foi encontrado no cache
9	L1 dcache Stores	# de veze que algo foi armazenado no dcache de nível 1
10	L1 dcache Loads	# de vezes que algo foi lido no dcache de nível 1
11	L1 dcache LoadMisses	# de vezes que alguma informação não foi encontrada no dcache de nível 1
12	LLC Stores	# de vezes que informação foi armazenada no cache de último nível
13	LLC Store Misses	# de vezes que houve falha de escrita no cache de último nível
14	LLC Loads	# de vezes que houve consulta ao cache de último nível
15	LLC Load Misses	# de vezes que algo não foi encontrado no cache de último nível
16	Page Faults	# de vezes que dados armazenados não foram encontrados na memória RAM física
17	Minor Faults	# de vezes que uma falta de página foi solucionada sem acessar o disco
18	Runtime	O tempo (em segundos) que levou para a aplicação finalizar sua execução
19	Thread	# de threads que a aplicação foi executada
20	Matrix Workload	Tamanho total do problema (matriz)
21	Energy Joules	O total de energia consumida por todos os componentes durante a execução da aplicação

Tabela 2: Contadores de desempenho utilizados neste trabalho

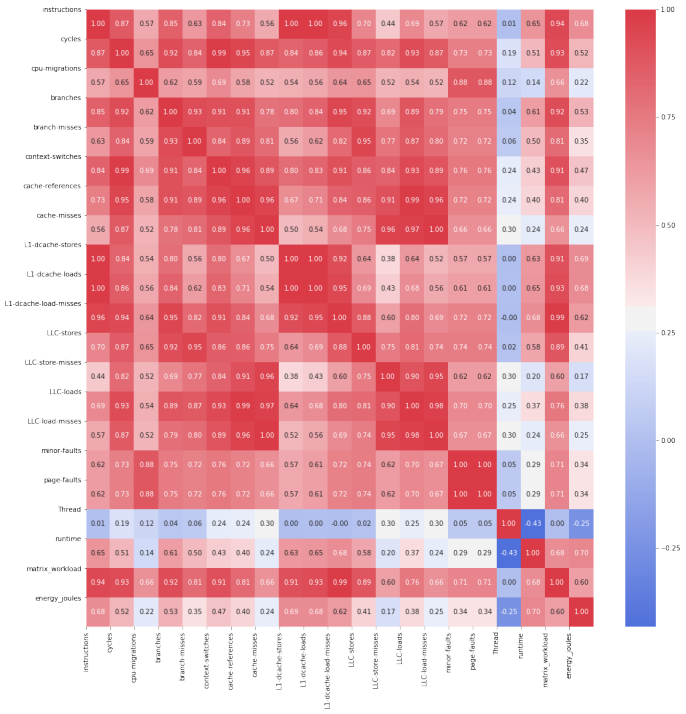


Figura 1: Mapa de calor gerado entre os atributos de aprendizado

máximo de 2.67GHz, duas placas de vídeo NVIDIA GF100GL GPUs e 24GB de memória RAM dividida em seis 6 pentes de 4096MB com velocidade máxima de 1333MHz. Esse ambiente também possuí três níveis de memória cache de tamanho 64KB, 256KB e 12288KB respectivamente.

O conjunto experimental, apresentado na Tabela 3, é composto por três aplicações (BT, LU e SP) que fazem parte do pacote *NAS Parallel Benchmark Suite*, sendo estas das classes Motif *Dense Linear Algebra* e *Structured Grids*.

	Tamanho da Matriz	# de Execuções	Classe Motif
BT A	64x64x64	210 (30x7)	Dense Linear Algebra
BT B	102x102x102	210 (30x7)	Dense Linear Algebra
BT C	162x162x162	210 (30x7)	Dense Linear Algebra
LU A	64x64x64	210 (30x7)	Dense Linear Algebra
LU B	102x102x102	210 (30x7)	Dense Linear Algebra
LU C	162x162x162	210 (30x7)	Dense Linear Algebra
SP A	64x64x64	210 (30x7)	Structured Grids
SP B	102x102x102	210 (30x7)	Structured Grids
SP C	162x162x162	210 (30x7)	Structured Grids

Tabela 3: Conjunto Experimental

O modelo de Árvore de Decisão Regressora desenvolvido se mostrou particularmente efetivo em predizer aplicações do mesmo tipo dos dados originais, mas não apresenta um resultado tão satisfatório em generalização em problemas muito diferentes dos dados originais usados para o treinamento. Na Tabela 4 são apresentadas as métricas de precisão associadas ao treinamento com os dados originais, após o balanceamento e seleção de parâmetros, e com o tempo de execução como o alvo da predição.

Erro Médio Absoluto	7.103270890466811
Erro Médio Quadrado	196.6353057785656
Porcentagem do Erro Médio Absoluto	12.019339%
Raiz Quadrada do Erro Médio	14.022671135649071
Coeficiente de Determinação (R2)	0.9965073159979737

Tabela 4: Precisão e Erro

Evidencia-se que o modelo possuí um alto coeficiente de determinação, mostrando uma alta capacidade de predizer novos dados e também uma porcentagem média de erro de aproximadamente 12,02% em relação ao nosso alvo.

5 Publicações

Durante o período de participação no PIBIC foram realizadas as seguintes submissões (ainda em avaliação):

[Gritz et al. 2019] - Gritz, M., Silva, G., Ferro, M., and Schulze, B. (2019). Towards an autonomous framework for hpc optimization: A study of performance prediction using hardware counters and machine learning. XIX Simpósio de Pesquisa Operacional e Logística da Marinha (SPOLM 2019). Sob avaliação - Submetido em Maio/2019.

[Silva et al. 2019] - Silva, G. D., Klôh, V. P., Yokoyama, A., Gritz, M., Ferro, M., and Schulze, B. (2019). SMCis: scientific applications monitoring and prediction for HPC environments. Springer. Trabalho selecionado entre os melhores do WSCAD 2018, convidado para a submissão de uma versão estendida para a Springer. (Submetido)

Neste momento encontra-se em preparação um artigo a ser submetido ao Workshop de Sistemas Computacionais de Alto Desempenho (WSCAD-WIC 2019).

6 Conclusões

Neste trabalho é apresentada a pesquisa em andamento para desenvolver um framework autonômico capaz de fazer a orquestração entre aplicações, escalonadores e arquiteturas, com base nos requisitos das aplicações científicas. Foi apresentada a forma de coleta dos parâmetros relevantes, buscando compreender as aplicações e seu desempenho para diferentes modelos (Classe de Motif [Asanovic et al. 2009]), com foco no desenvolvimento das técnicas preditivas que farão parte do framework. As tarefas preditivas estão sendo desenvolvidas

usando técnicas de aprendizado de máquina. Neste trabalho apresentamos todas as etapas para prever o tempo de execução de uma aplicação usando a técnica de árvore de regressão.

Os resultados de treinamento e teste tiveram um bom desempenho e também o processo de generalização parece bom ao observar a árvore e o número de exemplos cobertos por cada folha. Entretanto, o erro real ao se avaliar o modelo preditivo com novos casos nunca vistos foi muito alto. A especialista do domínio atribui essa baixa qualidade preditiva das regras a falta de atributos considerados representativos para esse domínio, os quais não haviam sido incluídos na tarefa de aprendizagem. Assim, o modelo preditivo gerado foi validado com novos exemplos e a suspeita foi confirmada com resultados insatisfatórios no processo de validação. Apesar dos resultados já terem uma precisão aceitável para este tipo de problema, ainda precisa ser melhorada. Além disso, os resultados ainda são iniciais e novos experimentos, com mais exemplos e em arquiteturas variadas precisam ser realizadas para confirmar a qualidade preditiva do modelo.

7 Agradecimentos

Agradeço ao Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq) pelo apoio financeiro e ao LNCC pela oportunidade concedida.

Referências

[Asanovic et al. 2009] Asanovic, K., Bodik, R., Demmel, J., Keaveny, T., Keutzer, K., Kubiatowicz, J., Morgan, N., Patterson, D., Sen, K., Wawrzynek, J., Wessel, D., and Yelick, K. (2009). A View of the Parallel Computing Landscape. *Commun. ACM*, 52(10):56–67. 5

[Gritz et al. 2019] Gritz, M., Silva, G., Ferro, M., and Schulze, B. (2019). Towards an autonomous framework for hpc optimization: A study of performance prediction using hardware counters and machine learning. *XIX Simpósio de Pesquisa Operacional e Logística da Marinha (SPOLM)*. 5

[HPC4e 2017] HPC4e (2017). High performance for energy (hpc4e). 1

[Kogge et al. 2008] Kogge, P., Bergman, K., Borkar, S., Campbell, D., Carlson, W., Dally, W., Denneau, M., Franzon, P., Harrod, W., Hill, K., Hiller, J., Karp, S., Keckler, S., Klein, D., Lucas, R., Richards, M., Scarpelli, A., Scott, S., Snavely, A., Sterling, T., Williams, R. S., and Yelick, K. (2008). ExaScale Computing Study: Technology Challenges in Achieving ExaScale Systems. Technical report, DARPA IPTO, Air Force Research Labs. 1

[Messina 2017] Messina, P. (2017). The exascale computing project. *Computing in Science Engineering*, 19(3):63–67. 1

[Patterson 2013] Patterson, D. (2013). *Orgins and Vision of the UC Berkeley Parallel Computing Laboratory*, chapter 1, pages 11–42. Microsoft Corporation, 1 edition. 1

[Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. 3

[Silva et al. 2019] Silva, G. D., Klôh, V. P., Yokoyama, A., Gritz, M., Ferro, M., and Schulze, B. (2019). *SMCis: scientific applications monitoring and prediction for HPC environments*. Springer. Trabalho selecionado entre os melhores do WSCAD 2018 convidado para a submissão de uma versão estendida para a Springer. 5

Título do projeto: A interação da cevada com os nutrientes do solo e o pulgão afídeo.
Nome do bolsista: Priscila Luana Lopes de Barros Weisz
Nome do orientador Maurício Vieira Kritz
Nome do Co orientador: Lucas dos Anjos
Tipo de Bolsa: Iniciação Científica
Período do relatório: Fevereiro a julho de 2019
Data: 13/07/2019

A INTERAÇÃO DA CEVADA COM OS NUTRIENTES DO SOLO E COM O AFÍDEO

Resumo: Este trabalho visa modelar o fenômeno, inseto/planta e compreender como ocorrem as interações entre o indivíduo vegetal (cevada), o afídeo (pulgão) e os nutrientes do solo que entrarão como o principal componente proporcionando diferentes respostas de cada cepa da cevada. .

Palavras-chave: Cevada – Afídeo – Nutrientes e Solo.

1) Introdução

De acordo com Ullmann (2002), a cevada é um cereal de inverno da família Gramineae, utilizado na indústria cervejeira (para a preparação do malte), na fabricação de rações, na indústria de farinha para alimentação infantil, na fabricação de doces e confeitos, na

panificação e ainda para fins terapêuticos. A espécie de maior importância é Hordeum vulgare L e será abordada neste trabalho.

Com base no artigo estudado, deve-se considerar também influência do solo, que fornece nutrientes para o desenvolvimento dos organismos destas populações envolvidas, visto que populações de plantas absorvem nutrientes a partir do solo.

De acordo com (Pereira, Lau, Júnior e Panizzi 2012), várias espécies de afídeos ou pulgões (*Hemiptera, Aphididae*), ocorrem na cultura de trigo, dependendo da época do ano e da região tritícola. As mais comuns são o pulgão-verde- dos-cereais, *Schizaphis graminum* (Rondani,1852); o pulgão-do-colmo-do trigo ou pulgão-da-aveia, *Rhopalosiphum padi* (Linnaeus, 1758);o pulgão-da-folha- do-trigo, *Metopolophium dirhodum* (Walker, 1849), e o pulgão-da-espiga-do trigo, *Sitobion avenae* (Fabricius, 1794), o qual será o foco de estudo deste trabalho.

Astles e Jones, apud (Rowntree , Vennon e Preziosi 2010), apontam que "os efeitos de interações genotípicas, podem ser contraditórios" e que o estado nutricional da planta pode influenciar a resposta do ataque pelos pulgões. Assim Rowntree, Vennon e Preziosi (2010) consideraram seis genótipos diferentes de cevada. Cada genótipo possui características diferentes no que se refere à absorção de nutrientes.

Este trabalho, pretende compreender o experimento descrito em (Rowntree ,Vennon e Preziosi 2010), feito em laboratório, compreendendo todas as etapas do fenômeno de suas interações . Durante a realização deste estudo, está sendo desenvolvido um mapa conceitual para facilitar a construção de um modelo matemático computacional buscando entendimento e visualização dos processos envolvidos.

2) Objetivos:

• O presente trabalho tem como objetivo avaliar, como ocorre a interação entre a cevada e o afídeo (pulgão), considerando a entrada de nutrientes do solo e a resposta que cada um dos seis genótipos de cevada irão gerar, através de determinadas quantidades absorvidas de nutrientes.

• Compreender o comportamento da interação entre inseto e planta, considerando a entrada de nutrientes mediada pelo solo como fonte de recurso da planta e a cevada como fonte de recurso do afídeo. Vale lembrar que a observação da capacidade de absorção também foi avaliada .

3) Métodos:

- Apresentar como foi avaliado na pesquisa , o processo de crescimento da planta. Neste caso, foi observado por meio da taxa de crescimento da cevada e sua interação com o pulgão afídeo dentro do artigo. Na pesquisa, será utilizado um modelo matemático para representar o que ocorrreu no experimento segundo o artigo.

- Elaborar um mapa conceitual a partir do artigo que ressalte os pontos principais do fenômeno, as hipóteses feitas e os métodos de observação.

- Elaboração de um modelo matemático computacional a partir do mapa conceitual que representa o fenômeno e respeite os princípios e comportamentos estudados em biologia teórica.

- A taxa de crescimento será usada como meio de avaliação do efeito dos genótipos da cevada e de suas capacidades de absorção de nutrientes na relação inseto- planta.

- Analisar quais foram os meios de entrada dos nutrientes através do solo em cada planta cultivada em vasos e suas diferentes proporções de nutrientes, para a observação do processo de interação comportamental dos afídeos e da planta cevada. Os genótipos trabalhados no artigo foram:G1Morex, G2Triumph, G3Blenheim, G4 Baronesa, G5 BCD47 e G6 Promessa Douradada. No experimento, foi observado de que forma os afídeos foram atraídos por cada cevada.

- Identificar no artigo, se existe uma correlação positiva do afídeo com a de cevada e observar o fator determinante para o crescimento da população de pulgões.

- Realizar uma comparação no genótipo que causou aumento no tamanho da planta. Relacionar, com o aumento da concentração de nutrientes com os genótipos que cresceram mais rápido e comparar com a diminuição da concentração de nutrientes.

4) Estado Atual:

A partir do estudo de (Rowntree, Vennon e Preziosi 2010), está sendo elaborado um mapa conceitual descrevendo o mesmo. Este mapa propiciará os meios para o desenvolvimento do modelo e sua utilização no estudo do fenômeno.

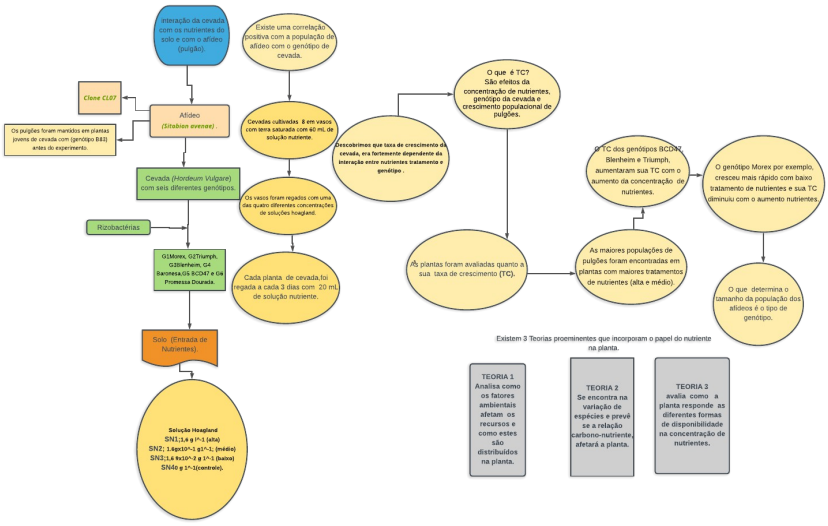


Figura 1: Estado atual do mapa conceitual baseado em (Rowntree J;Vennon. A; Preziosi R.2010).

A partir da conclusão do mapa conceitual, será elaborado um modelo matemático, computacional para avaliar através de experimentos numéricos os seguintes comportamentos populacionais:

1) O Crescimento exponencial da planta e resposta funcional do tipo 1 do afídeo:

2) O Crescimento logístico da planta e resposta funcional do tipo 1 do afídeo:

3) O Crescimento exponencial da planta e resposta funcional do tipo 2 do afídeo:

4) O Crescimento logístico da planta e resposta funcional do tipo 2 do afídeo.

5) Resultados esperados

Através do modelo matemático, computacional, simples e baseado no de Lotka -Volterra é esperado entender de forma mais clara e precisa, o fenômeno descrito no artigo.

É importante salientar que a partir do trabalho realizado no presente projeto haverá possibilidade de haver uma compreensão adicional sobre o que aconteceu no experimento, de forma a explicar sob a ótica da ecologia teórica a relação de interação entre os componentes do sistema.

6) Conclusão

Este trabalho, teve como base de estudo, o artigo descrito por (Rowntree J; Vennon. A; Preziosi R. 2010), e a partir do mesmo,está sendo construido um mapa conceitual. Este mapa visa atuar também como ferramenta para elaboração de experimentos numéricos através de modelos matemáticos computacionais.

Referências Bibliográficas:

Astles et al (2005) Genetic variation in response to an indirect ecological effect. Proc R Soc Lond B 272:2577–2581 Apud (Rowntree J;Vennon. A; Preziosi R.2010).

Te ́tard-Jones (2007) Genotype-by-genotype interactions modified by a third species in a plant–insect system. Am Nat 170:492–499 Apud (Rowntree J;Vennon. A; Preziosi R.2010).

Rowntree J;Vennon. A; Preziosi R.(2010) Plant genotype mediates the effects of nutrients on aphids. Oecologia (2010), 163.675-679,DOI 10.1007//s00442-010-1609-1.

PEREIRA, Paulo. LAU, Douglas. JÚNIOR, Alberto. PANIZZI, Antonio. Fitossanidade. Pragas iniciais que comprometem até o fim. Revista A Granja. Editora Centaurus. Porto Alegre – RS. Edição 763.Publicada em julho de 2012. Disponível em https://edcentaurus.com.br/agranja/edicao/763/materia/4481. Acesso em 15/07/2019.

ULLMANN Samanta (2002), A feira. Cereais. Cevada. Porto Alegre- RS abril/2002. Disponível em https://plone.ufrgs.br/afeira/materias-primas/cereais/cevada acesso em 16/07/2019.

Avaliação de Métodos de Aprendizado em Aplicações de Saúde

Nome do Bolsista: Raquel de Abreu Junqueira Gritz
Nome do Orientador: Fábio André Machado Porto
Nome do Coorientador: Douglas Ericson Marcelino de Oliveira
Tipo de Bolsa e Período do Relatório: Bolsa de Iniciação Científica – IC / PIBIC – Julho de 2019

1. Objetivos

Este trabalho de iniciação científica tem como objetivo utilizar técnicas de pré-processamento de dados de saúde que influenciarão no treinamento e serão úteis para determinar uma predição com qualidade, visto que tomadas de decisão no âmbito da saúde não podem conter falhas.

Tendo em vista a busca por uma predição de qualidade, busca-se também comparar as técnicas de aprendizado em Árvore de Decisão e Random Forest aplicadas a problemas ligados à predição, em particular considerando dados de saúde relacionados à causas de óbito, no contexto da cooperação LNCC-DEXL / FIOCRUZ-ICICT, e identificar características dos métodos que melhor se adéquem ao problema.

2. Introdução

A busca pela capacidade de analisar um grande volume de dados por meio de métodos estatísticos e o uso de uma variedade de algoritmos para encontrar padrões nos dados, mostra que, com base nesses padrões, é possível realizar predições.

Uma área em que predições são bem vindas é a saúde. Atualmente, um grande número de dados encontra-se disponível e que pode ser usado em um processo de predição. Neste contexto, este trabalho utilizou dados referentes à causas de óbito em todos os estados brasileiros ao longo dos anos de 1996 a 2014. O alvo utilizado no modelo de predição foi o atributo que identifica causas de morte relacionadas a doenças do aparelho circulatório e doenças do aparelho respiratório.

O projeto deu início com o estudo dos dados, buscando encontrar padrões ao longo dos anos em todos os estados. Nele foi possível identificar mudanças em atributos. A partir disso foi realizada uma seleção de atributos de treinamento que fossem relevantes ao alvo e que pudessem influenciar a performance do modelo positivamente. Com isso 40 atributos de um total de 156 foram selecionados para o conjunto de treinamento. No atributo alvo as classes de dados que definiam outros tipos de causas de óbito que não fossem relacionadas à doenças do aparelho circulatório e respiratório foram retiradas.

Técnicas de pré-processamento de dados foram utilizadas onde, através da limpeza, foi possível identificar valores faltantes e assim preparar os dados para a etapa de treinamento. Foi utilizada a técnica de aprendizagem supervisionada baseada em Árvore de Decisão, assim como Random Forest. Os dados foram divididos em conjuntos de treino e teste, estes foram treinados e foi avaliada a qualidade da predição.

Na continuação deste trabalho, pretendemos escrever um artigo sobre os resultados obtidos, assim como avaliar a técnica de Árvore de Decisão sobre dados de saúde do DATASUS sobre Hospitalizações, com o objetivo de predizer o estado final de um paciente quando der entrada em uma internação.

3. Metodologia

Esse trabalho considerou dados de estados brasileiros relacionados à causas de óbito ao longo dos anos de 1996 a 2014. Esses dados estavam divididos em diversos arquivos que correspondiam a um estado em determinado ano. Ao longo da etapa de pré-processamento foram se encontrando dificuldades que precisaram ser solucionadas. O trabalho teve como tarefas:

a) observar os conjuntos de dados separados por estados para identificar se os padrões se seguiam ao longo dos anos. Ao se deparar com mudanças nos atributos ao longo dos anos, buscou-se identificar se todos os estados seguiam essa mesma característica;

b) criação de um algoritmo para determinar a quantidade de atributos nos conjuntos de dados correspondentes a determinado estado ao longo dos anos. Este conseguiu determinar quais atributos foram incluídos e/ou excluídos no decorrer dos anos;

c) seleção de atributos relevantes para o alvo;

d) descoberta de padrões nos atributos selecionados em um conjunto de anos, que foram denominados blocos;

e) escolha do alvo para treinamento que corresponde ao atributo relacionado à causas de óbito;

f) retirada de classes do atributo alvo que identificavam outras causas de óbito que não fossem das classes de doenças do aparelho circulatório e doenças do aparelho respiratório;

g) identificação de valores faltantes nos conjuntos de dados e preenchimento dos mesmos;

h) em uma primeira etapa, realização de treinamento e teste com Árvore de Decisão em cada um dos blocos em separado;

i) em seguida, realização de treinamento e teste com Árvore de Decisão no conjunto de dados de todos os estados e todos os anos;

j) finalmente, realização de treinamento e teste com Árvore de Decisão em cada um dos estados em separado;

k) busca de padrões entre estados e o conjunto total de anos e estados;

l) devido ao retorno de uma baixa predição e uma alta taxa de erro em dados de doenças do aparelho respiratório, buscamos trabalhar o balanceamento do alvo juntamente com o uso de outro algoritmo denominado Random Forest.

4. Resultados e Discussão

Neste trabalho foram usados os algoritmos de Árvore de Decisão e Random Forest em Python com Scikit-learning tendo como base dados relacionados a causas de óbito nos estados brasileiros entre os anos de 1996 e 2014.

Teve como objetivo retornar um modelo em Árvore de Decisão e Random Forest de boa precisão que determinaria as causas de óbito relacionadas a doenças do aparelho circulatório e doenças do aparelho respiratório.

4.1. Conjunto de Dados Desbalanceado

O conjunto de dados usado para treinamento e avaliação é composto por dados de todos os anos (1996 a 2014) e estados brasileiros. Este conjunto de dados possui 41 atributos, sendo um deles usado como atributo alvo. Foi possível retornar a matriz de confusão dos modelos em Árvore de Decisão e Random Forest. O F1-score do modelo em Árvore de Decisão obteve aproximadamente 0.85 para valores relacionados a doenças do aparelho circulatório, enquanto que 0.04 para valores relacionados a doenças do aparelho respiratório. Já o F1-score do modelo em Random Forest obteve aproximadamente 0.86

para valores relacionados a doenças do aparelho circulatório e 0.05 para valores relacionados a doenças do aparelho respiratório.

Na Tabela 1 é possível visualizar a matriz de confusão do modelo em Árvore de Decisão onde, dos 935776 valores determinados como de doenças cardíacas, o algoritmo identificou 934783 de forma correta, enquanto que dos 334948 de valores determinados como de doenças respiratórias, ele acertou somente 6903. Através da tabela abaixo é possível visualizar que os valores relacionados à doenças do aparelho respiratório possuem uma alta taxa de erro em relação a valores de doenças do aparelho circulatório.

Valor Predito	0	1	All
Valor Real			
0	934783	993	935776
1	328045	6903	334948
All	1262828	7896	1270724

0 - Doenças do aparelho circulatório
1 - Doenças do aparelho respiratório

Tabela 1 – Matriz de Confusão da Árvore de Decisão de Todos os Estados e Anos

Na Tabela 2 é possível visualizar a matriz de confusão do modelo em Random Forest onde, dos 935776 valores determinados como de doenças cardíacas, o algoritmo identificou 934696 de forma correta, enquanto que dos 334948 de valores determinados como de doenças respiratórias, ele acertou somente 8225. O modelo em Random Forest possui uma taxa de acerto em relação a valores de doenças respiratórias um pouco maior do que o modelo em Árvore de Decisão. No entanto, a taxa de erros ainda é alta em comparação à doenças circulatórias.

Valor Predito	0	1	All
Valor Real			
0	934696	1080	935776
1	326723	8225	334948
All	1261419	9305	1270724

0 - Doenças do aparelho circulatório
1 - Doenças do aparelho respiratório

Tabela 2 – Matriz de Confusão da Random Forest de Todos os Estados e Anos

Um outro experimento foi realizado com o conjunto de dados de 1996 a 2014 do estado do Rio de Janeiro. Este conjunto de dados possui 41 atributos, sendo um deles usado como atributo alvo e foi possível retornar a matriz de confusão dos modelos em Árvore de Decisão e Random Forest. O F1-score do modelo em Árvore de Decisão obteve aproximadamente 0.86 para valores relacionados a doenças do aparelho circulatório e 0.22 para valores relacionados a doenças do aparelho respiratório. Já o F1-score do modelo em Random Forest obteve aproximadamente 0.86 para valores relacionados a doenças do aparelho circulatório e 0.24 para valores relacionados a doenças do aparelho respiratório.

Na Tabela 3 é possível visualizar a matriz de confusão do modelo em Árvore de Decisão onde, dos 133459 valores determinados como de doenças cardíacas, o algoritmo identificou 133459 de forma correta, enquanto que dos 49224 de valores determinados como de doenças respiratórias, ele acertou 6092. Assim como o conjunto de dados de todos os estados e anos, o estado do Rio de Janeiro também possui uma alta taxa de erro nos valores relacionados à doenças do aparelho respiratório no modelo em Árvore de Decisão.

Valor Predito	0	1	All
Valor Real			
0	133459	0	133459
1	43132	6092	49224
All	176591	6092	182683

0 - Doenças do aparelho circulatório
1 - Doenças do aparelho respiratório

Tabela 3 – Matriz de Confusão da Árvore de Decisão do Estado do Rio de Janeiro

Já na Tabela 4 é possível visualizar a matriz de confusão do modelo em Random Forest onde, dos 133459 valores determinados como de doenças cardíacas, o algoritmo identificou 133381 de forma correta, enquanto que dos 49224 de valores determinados como de doenças respiratórias, ele acertou 6878. O modelo em Random Forest possui uma alta taxa de erro em relação a valores de doenças do aparelho respiratório da mesma forma como visto acima no modelo em Árvore de Decisão.

Valor Predito	0	1	All
Valor Real			
0	133381	78	133459
1	42346	6878	49224
All	175727	6956	182683

0 - Doenças do aparelho circulatório
1 - Doenças do aparelho respiratório

Tabela 4 – Matriz de Confusão da Random Forest do Estado do Rio de Janeiro

4.2. Conjunto de Dados Balanceado

No entanto, devido a baixa taxa de acertos em relação aos valores relacionados a doenças do aparelho respiratório, a solução encontrada foi criar uma função que realizasse a técnica de *oversampling* - sobre amostragem aleatória, onde foi possível replicar instâncias da classe minoritária e assim balancear os dados do atributo alvo em relação aos exemplos da classe de doenças circulatórias. Em vista disso, foi possível, nesta fase inicial, fazer com que a taxa de erros em relação aos dados de doenças do aparelho respiratório diminuíssem.

O conjunto de dados composto por todos os estados em todos os anos após o balanceamento obteve o F1-score do modelo em Árvore de Decisão de 0.58 para valores relacionados a doenças do aparelho circulatório e 0.68 para valores relacionados a doenças do aparelho respiratório. Já o F1-score do modelo em Random Forest obteve aproximadamente 0.62 para valores relacionados a doenças do aparelho circulatório e 0.64 para valores relacionados a doenças do aparelho respiratório.

Na Tabela 5 é possível visualizar a matriz de confusão do modelo em Árvore de Decisão de Decisão onde, dos 597663 valores classificados como de doenças do aparelho circulatório, o algoritmo pode identificar 199990 valores de forma correta, enquanto que dos 596724 valores classificados como de doenças do aparelho respiratório, o acerto foi de 512603 valores. Com os valores balanceados a taxa de acertos de doenças respiratórias no modelo em árvore de decisão melhora consideravelmente, mas ainda apresenta dificuldades em classificar os valores de forma correta nas duas classes.

Valor Predito	0	1	All
Valor Real			
0	199990	397673	597663
1	84121	512603	596724
All	284111	910276	1194387

0 - Doenças do aparelho circulatório
1 - Doenças do aparelho respiratório

Tabela 5 – Matriz de Confusão da Árvore de Decisão de Todos os Estados e Anos Balanceados

A Tabela 6 demostra a matriz de confusão do modelo em Random Forest onde, dos 597663 valores determinados como doenças do aparelho circulatório, 361912 valores foram identificados pelo algoritmo de forma correta, enquanto que dos 596724 valores determinados como doenças do aparelho respiratório, o algoritmo acertou 392517 valores. O modelo em Random Forest apresenta uma taxa de acertos melhorada na classificação de doenças respiratórias após o balanceamento, mas ainda apresenta dificuldades em realizar a classificação corretamente.

Valor Predito	0	1	All
Valor Real			
0	361912	235751	597663
1	204207	392517	596724
All	566119	628268	1194387

0 - Doenças do aparelho circulatório
1 - Doenças do aparelho respiratório

Tabela 6 – Matriz de Confusão da Random Forest de Todos os Estados e Anos Balanceados

Em relação ao conjunto de dados do estado do Rio de Janeiro após o balanceamento, o F1-score do modelo em Árvore de Decisão obteve aproximadamente 0.42 para valores relacionados a doenças do aparelho circulatório e 0.69 para valores relacionados a doenças do aparelho respiratório. Já o F1-score do modelo em Random Forest obteve aproximadamente 0.78 para valores relacionados a doenças do aparelho circulatório e 0.81 para valores relacionados a doenças do aparelho respiratório.

A tabela 7 demonstra a matriz de confusão do modelo em Árvore de Decisão onde, dos 133330 valores determinados como de doenças cardíacas, o algoritmo identificou 39193 de forma correta, enquanto que dos 133380 de valores determinados como de doenças respiratórias, ele acertou 121116. No modelo em Árvore de decisão com dados balanceados a taxa de acerto em relação à doenças do aparelho respiratório aumenta em comparação com valores não balanceados e ultrapassa a taxa de erro.

Valor Predito	0	1	All
Valor Real			
0	39193	94137	133330
1	12264	121116	133380
All	51457	215253	266710

0 - Doenças do aparelho circulatório
1 - Doenças do aparelho respiratório

Tabela 7 – Matriz de Confusão da Árvore de Decisão do Estado do Rio de Janeiro Balanceado

E na Tabela 8 é possível visualizar a matriz de confusão do modelo em Random Forest onde, dos 133330 valores determinados como de doenças cardíacas, o algoritmo identificou 98634 de forma correta, enquanto que dos 133380 de valores determinados como de doenças respiratórias, ele acertou 113818. A taxa de acerto dos valores relacionados à doenças do aparelho respiratório no modelo em Random Forest é alta em comparação com o mesmo modelo que usa valores desbalanceados. Isso significa que ao balancear os valores de doenças respiratórias a taxa de erro diminui.

Valor Predito	0	1	All
Valor Real			
0	98634	34696	133330
1	19562	113818	133380
All	118196	148514	266710

0 - Doenças do aparelho circulatório
1 - Doenças do aparelho respiratório

Tabela 8 – Matriz de Confusão da Random Forest do Estado do Rio de Janeiro Balanceado

Foi possível, assim, chegar ao que era pretendido inicialmente, compreendendo a importância da etapa de pré-processamento dos dados tendo como resultado uma precisão de qualidade frente aos desafios enfrentados. Como visto através das tabelas a classificação em doenças do aparelho respiratório tem uma melhora significativa com o balanceamento de dados, enquanto que as doenças circulatórias tem uma leve piora em sua classificação quando as classes se apresentam balanceadas. Ainda não temos uma resposta definitiva para este problema, no entanto, supomos que isso ocorra devido as duas classes que, por estarem intercaladas em relação aos atributos de treinamento, não conseguem determinar claramente quais são as classes reais, como pode ser visto na Figura 1 onde a curva de distribuição das duas classes de causas de óbito se intercala em relação a um dos atributos - OCUP.

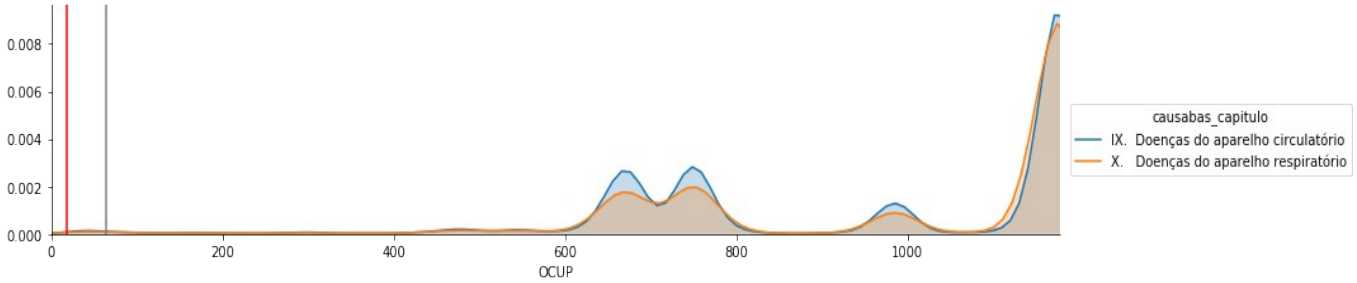


Figura 1 – Gráfico de Curva de Distribuição de Valores

No anexo I é possível visualizar as imagens das Árvores de Decisão dos resultados acima descritos. Os atributos presentes correspondem a *def_loc_ocor* que define o local de ocorrência do óbito (hospital, outros estabelecimentos de saúde, domicílio, via pública); *def_parto* que define o tipo de parto relacionado à causa de óbito; *def_escol* que define os anos de escolaridade do falecido; *def_est_civil* que define o estado civil do falecido; *ano_obito* que define o ano de óbito; OCUP que define a ocupação habitual e o ramo de atividade do falecido e NUMEROLOTE que define o número do lote na declaração de óbito.

5. Conclusão

Neste trabalho foi possível criar um modelo de aprendizado baseado na técnicas de Árvore de Decisão e Random Forest para classificação de atributos ligados à Causas de Óbito. Estes forma capazes de classificar novas entradas com características ligadas à causas de morte por doenças do aparelho circulatório e doenças do aparelho respiratório. Nas tabelas acima, os valores relacionados à doenças do aparelho respiratório possuem uma alta taxa de erro em comparação com valores relacionados à doenças do aparelho circulatório. Segundo o gráfico de curva de distribuição de valores dos atributos em relação ao alvo que foi visualizado neste trabalho na Figura 1, isso acontece porque os valores do atributo alvo estão intercalados entre si e não há uma divisão clara em relação a eles.

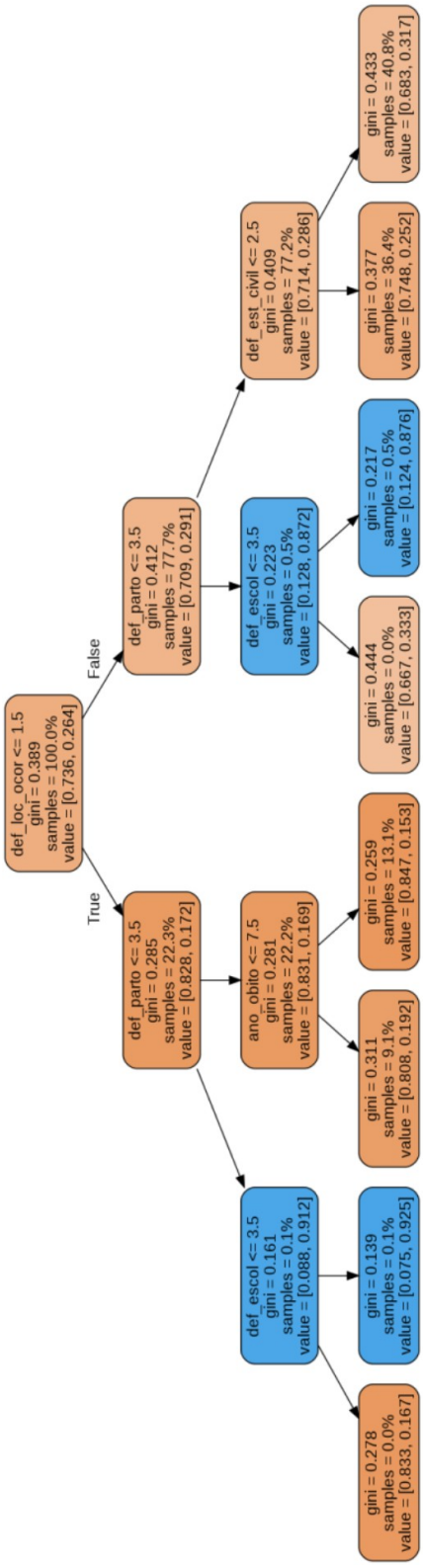
Como objetivo futuro buscaremos publicar um artigo sobre este trabalho e seus resultados baseados nas técnicas utilizadas ao longo do trabalho frente aos problemas apresentados. Temos também como objetivo futuro, aprimorar os estudos sobre o trabalho em realização e começar outro trabalho relacionado a dados de hospitalização para fazer a análise de dados para saúde onde se buscará fazer a predição determinando características que possam levar a alta de um paciente ou ao óbito do mesmo.

6. Referências Bibliográficas

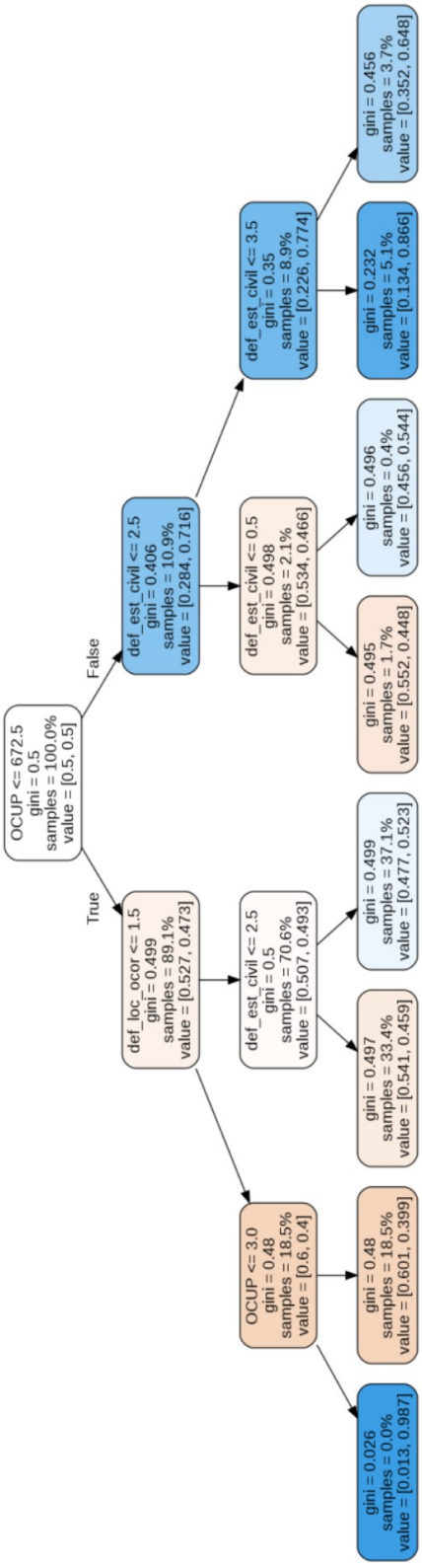
Thomas G. Dietterich, Ensemble Methods in Machine Learning, ISBN: 3-540-67704-6, Springer-Verlag London, UK, 2000

ANEXO I

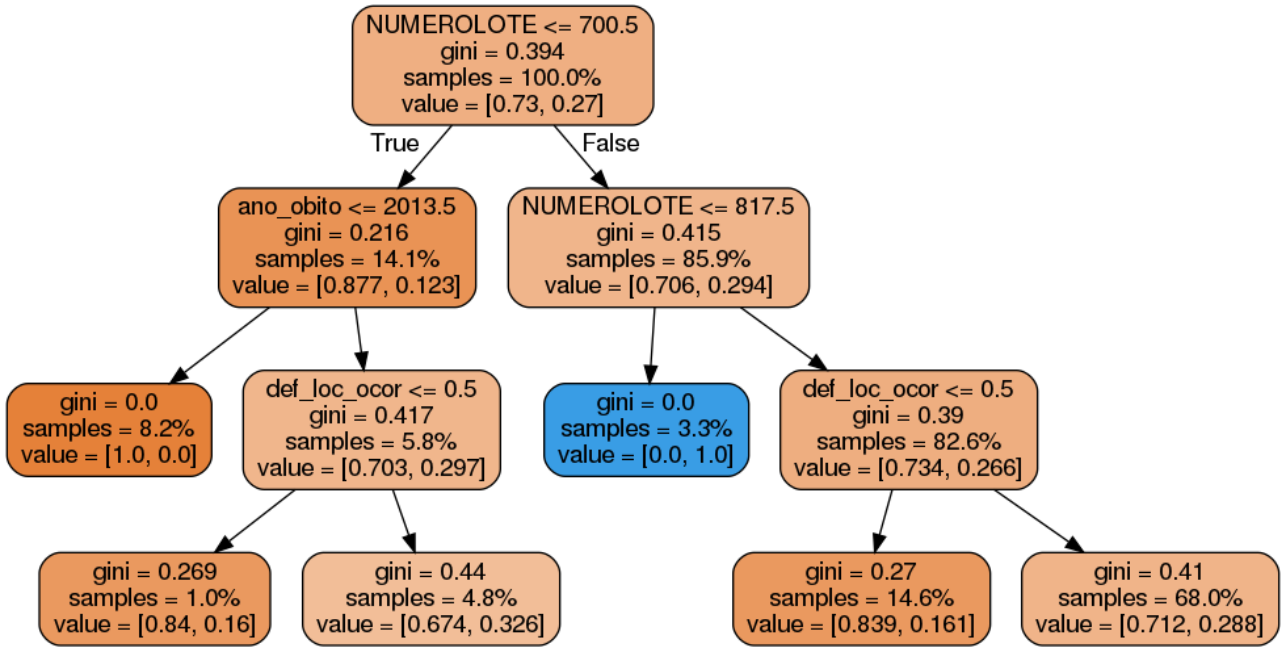
Árvore de Decisão de Todos os Estados e Anos



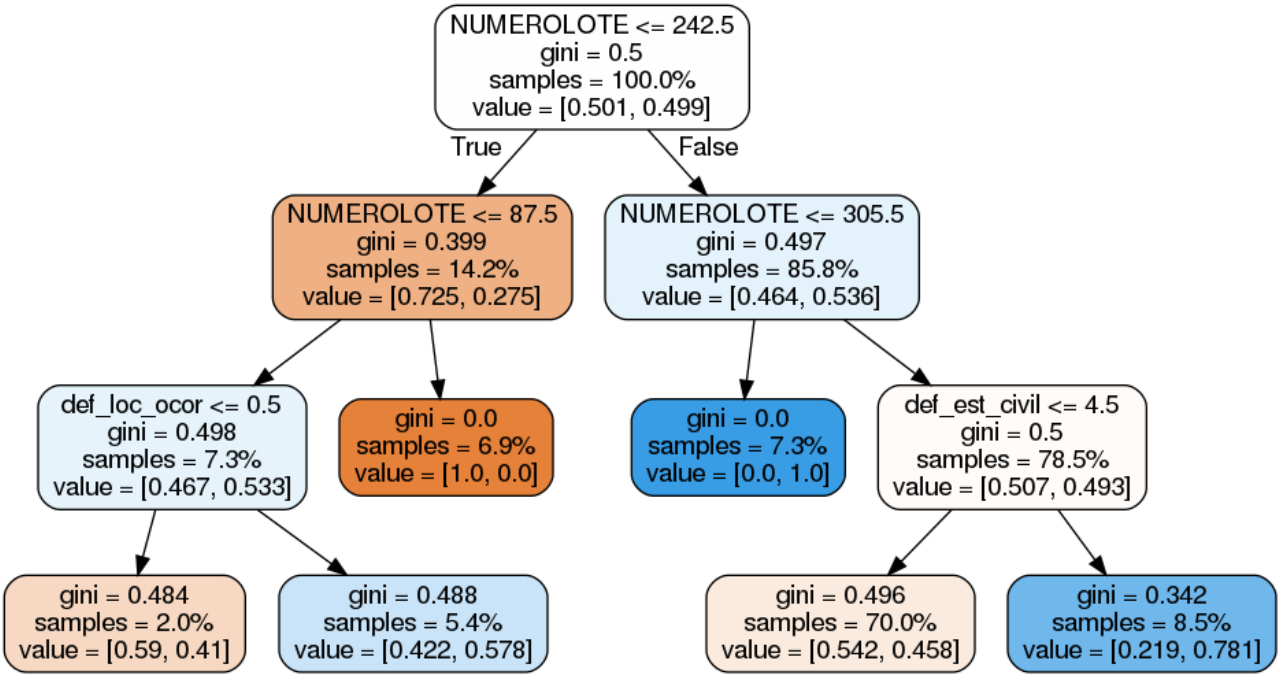
Árvore de Decisão de Todos os Estados e Anos com Balanceamento



Árvore de Decisão do Estado do Rio de Janeiro



Árvore de Decisão do Estado do Rio de Janeiro com Balanceamento



Bolsista:

René Constancio Nunes de Lima

Orientadores:

Elson M. Toledo(LNCC)

Regina Célia P. Leal Toledo

(UFF- co-orientadora)

Relatório- do projeto:

SIMULAÇÃO NUMÉRICA E COMPUTACIONAL DO TRÁFEGO VIÁRIO

LNCC

PROGRAMA PIBIC/PIBITI

2018-2019

1- Objetivos

O projeto em desenvolvimento tem o objetivo de analisar o tráfego viário utilizando uma modelagem microscópica, baseada em Autômatos Celulares (AC), considerando, em todas as análises, a influência de diferentes políticas de aceleração e desaceleração, principal característica do projeto em questão.

Este projeto dá continuidade ao trabalho que vinha sendo realizado pelo bolsista PIBIC/LNCC anterior, onde o tradicional modelo probabilístico de AC proposto por Nagel e Schreckenberg [1], conhecido como modelo NaSch, foi modificado para possibilitar a representação de diferentes comportamentos dos motoristas, aqui modelada por tendências, de cada perfil de motorista, para acelerar e desacelerar a cada instante de tempo. Na proposta apresentada, essa modificação consiste em utilizar uma discretização do espaço mais refinada do que a utilizada normalmente no modelo NaSch e uma Função de Densidade de Probabilidade (FDP) não uniforme. A FDP Beta foi utilizada para definir esses diferentes perfis de aceleração e desaceleração. Eles variam de um perfil mais cauteloso, que tem a tendência de acelerar/desacelerar mais suavemente, a um mais agressivo, que acelera e desacelera mais bruscamente, e são obtidas através das diferentes médias e variâncias da FDP utilizada.

Na etapa atual consideramos a influência desses perfis na dinâmica de pistas com cruzamentos. Esse cruzamento foi modelado como proposto em Marzoug [2], pretendendo avaliar como essas diferentes políticas podem influenciar nessa modelagem. Este relatório pretende mostrar alguns dos resultados já obtidos, bem como apresentar o que está em andamento no momento atual do projeto.

2- Introdução

O tráfego viário é um dos fatores que afeta diretamente a qualidade de vida das pessoas, particularmente quando pensamos no meio urbano. Várias soluções já foram propostas para tentar mitigar os efeitos do crescimento do número de veículos em grandes cidades, seja com veículos elétricos para diminuir a poluição do ar, ou veículos autônomos, que prometem reduzir os congestionamentos e o número de acidentes. Nesse contexto, analisar a dinâmica do tráfego, em diversas situações, é de extrema importância. Para fazer essas análises são utilizados modelos matemáticos que podem ser tanto macroscópicos quanto microscópicos. O modelo microscópico de Autômato Celulares (AC), foco deste trabalho, tem se mostrado eficiente e utilizado, com sucesso, em várias simulações. Uma das suas principais vantagens é que é facilmente implementado, tem baixa custo computacional e representa bem as características fundamentais do tráfego. No entanto, de forma geral, esses modelos, apesar de probabilísticos, tratam de forma uniforme todos os motoristas e, em algumas vezes, para diferenciá-los, variam apenas suas velocidades máximas. A pesquisa na qual esse projeto se insere pretende propor e avaliar a influência de diferentes comportamentos de motoristas na dinâmica do tráfego, modelando esses comportamentos por diferentes políticas de aceleração e frenagem. Modelos de AC em que as diferentes políticas de aceleração são consideradas foram inicialmente propostos em Zamith [3] e Zamith et al [4]. Essas políticas de aceleração são modeladas utilizando-se uma Função de Densidade de Probabilidade (FDP) não uniforme, a FPD Beta. Então, diferentes parâmetros da FDP Beta vão definir diferentes tendências de acelerações e frenagens, onde cada perfil de motorista tende a acelerar mais agressivamente ou cautelosamente. Também com esse enfoque, em etapas anteriores do projeto, o modelo NaSch tradicional foi modificado para avaliar os efeitos desses diferentes perfis na ocorrência de

situações perigosas, em pistas simples, e que podem levar a acidentes em vias expressas [5,6]. Na continuidade desse trabalho, que faz parte do trabalho realizado durante o período aqui considerado, estudamos formas para tratar pistas com cruzamento [2,7,8] e formas para considerar esses diferentes perfis desse tipo de modelagem. Para isso, utilizamos uma interseção em um circuito fechado, como proposto em Marzoug [2], onde duas pistas se cruzam perpendicularmente no meio. Marzoug mostrou que o diagrama fundamental depende fortemente da probabilidade de prioridade (P) no cruzamento e exibe quatro frases: fluxo livre, platô, congestionamento e uma nova fase ocorrendo para qualquer valor de $P \neq 0.5$. Essas fases desaparecem gradualmente a medida que P aumenta, e desaparecem completamente para P = 0.5. Aqui será mostrado como os diferentes perfis de aceleração, bem como diferentes probabilidades de frenagem alteram esses resultados.

Como o modelo NaSch é um modelo explícito, que não considera o movimento do veículo à frente no instante de tempo atual, apresentamos também, resumidamente, o andamento atual do projeto onde propomos um modelo de antecipação que tenta considerar, de alguma forma, esse movimento.

3- Material e Métodos ou Metodologia

Ao iniciar a vigência do presente projeto, já tínhamos conhecimento básico tanto do problema a ser tratado quanto de modelos simples de AC, como o modelo NaSch, pois tínhamos entrado neste grupo de pesquisa três meses antes com bolsa PIBIC/LNCC. Dando sequência, estudamos publicações [2,7] que modelavam um cruzamento simples e, nessa etapa, já interagi com a equipe em modificações do software CATS (Cellular Automata Traffic Simulation), que vem sendo desenvolvido pelo grupo. Dessa forma, participei da implementação do modelo proposto por Marzoug [2] modificado para simular os diferentes comportamentos de condutores.

Neste relatório apresentamos brevemente o modelo NaSch, e o modelo NaSch modificado. Depois, o modelo de interseção proposto por Marzoug [2], para então apresentar os resultados obtidos. Por fim, é descrita minha participação no andamento atual do projeto, onde estou implementando e testando políticas de antecipação, para integrar com o restante do grupo que está trabalhando em cruzamento de pistas e pistas múltiplas, em modelos mais representativos e complexos do que o sugerido por Marzoug, a fim de termos um modelo do tráfego mais realista.

Finalmente, cabe ressaltar que com essa etapa da pesquisa foi submetido e aceito o trabalho, do qual participei: “Influence of Drivers’ Behavior on Traffic Flow at Two Roads Intersection”, no 19th International Conference on Computational Science and Applications (ICCSA 2019), ocorrido de 1 – 4 de junho de 2019 in Saint Petersburg, Russia.

3.1 Modelo NaSch tradicional e NaSch modificado

(i) O modelo NaSch tradicional:

Nesse modelo de AC, em sua proposta tradicional, a via é unidimensional, com condição de contorno periódica e discretizada em células, onde cada célula ou é ocupada por um veículo ou está vazia. Cada veículo ocupa uma célula, normalmente considerada com *7,5m*, e o tempo é discretizado em instantes de tempo *t* de *1 segundo*. Cada veículo ocupa a posição *i*, no instante de tempo t, *x(i, t),* e tem velocidade dada por *v(i, t)*. A distância entre um veículo e outro, dada pelo número de células vazias na frente de cada veículo, é definida como $d(i, t) = x(i+1,t) - x(i,t) - L(i)$, sendo $L(i)$ o número de células ocupadas por um veículo *i*, tradicionalmente

considerado como *L(i) = 1* para todos os veículos *i*. O algoritmo consiste em um conjunto de quatro regras, aplicadas paralelamente a todos os veículos, a cada segundo, apresentado no Algoritmo 1.

Algorítmo 1

(1) Aceleração:	$v(i,t+1) = min[v(i,t) + A, V_{max}]$
(2) Desaceleração:	$v(i,t+1) = min[v(i,t+1), d(i,t)]$
(3) Desaceleração:	$v(i,t+1) = max[v(i,t+1) – A, 0]$, *com probabilidade* p_b
(4) Movimento:	$x(i,t+1) = x(i,t) + v(i,t+1)$

sendo V_{max} a velocidade máxima da via, *A = 1* a aceleração dos veículos e p_b, que modela a incerteza no comportamento dos motoristas, a probabilidade de o motorista não querer acelerar, ou desacelerar, dependendo da situação.

(ii) O modelo NaSch modificado:

No modelo NaSch original, *A = 1* e cada veículo acelera ou desacelera sempre de $1cel/s^2$, o que corresponde a $7{,}5m/s^2$. Para gerar diferentes políticas de aceleração, dividimos a célula de *7,5m* do NaSch em células menores, para que seja possível que cada motorista acelere, caso possa, de diferentes formas. Assim, o valor da variável *A,* no *Algoritmo 1,* é probabilístico e é definido, para cada motorista, pela FDP Beta, com diferentes médias e variâncias. Essa FDP gera um valor de α (0 < α < 1), que é utilizada para calcular a nova aceleração, dada por $A = int\ [\ (1-\alpha)\ A_{max}]$, onde int é uma função que retorna o inteiro mais próximo. Assim, o parâmetro *p* designa se o motorista vai acelerar ou não, e *A* como ele irá acelerar: mais suavemente ou mais bruscamente. A FDP Beta utilizada é definida como:

$$\beta(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, x^{a-1}(1-x)^{b-1}$$

onde $0 \leq x \leq 1$ *e* $\Gamma(n+1) = n!$, *n* um inteiro positivo. Dados os parâmetros *a* e *b*, resulta uma função $\beta(a,b)$. Para a geração desses valores utiliza-se a Técnica de Rejeição do Método de Monte Carlo. Valores de α próximos a *0* produzem acelerações mais próximas de A_{max}, enquanto valores próximos a *1* produzem acelerações próximas de *0*. Para cada par de valor *a* e *b*, podemos calcular a média $M = \frac{a}{a+b}$. Então, podemos relacionar cada perfil de aceleração a um par de parâmetros *a* e *b*, em que o valor médio determina a tendência de aceleração de um perfil. Para exemplificar, se a via for discretizada em células de *1,5 m*, é possível modelar o motorista que tende a acelerar/desacelerar mais suavemente, com média de *A = 1*, até o que tende a acelerar/desacelerar mais bruscamente, com média de *A = 4*. E quando *A* é sempre igual a *5* temos o modelo NaSch, com um veículo ocupando *5* células.

3.2 O modelo de interseção

O modelo de interseção proposto em Marzoug [2] foi implementado inicialmente pois ele modela o cruzamento em uma via simples, definida como um “oito”, como apresentado na Figura 1. Este modelo considera duas pistas, R1 e R2, perpendiculares e cruzando no meio. Em R1 os veículos se movem de cima para baixo e em R2, da esquerda para direita. Nesse circuito fechado, a saída de R1 é a entrada de R2, e a saída de R2 é a entrada de R1, como ilustra a Figura 1.

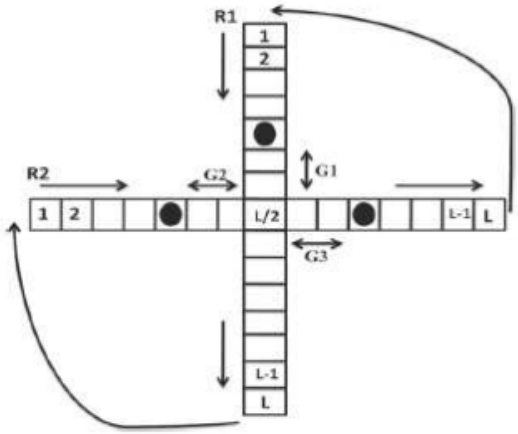


Figura 1

No modelo NaSch tradicional a interseção é composta por apenas uma célula. Na modificação proposta ela é composta pelo número de células que um veículo ocupa. Isso afeta o comportamento no cruzamento pois um veículo, dependendo do comportamento do motorista e do tráfego, pode demorar mais de um instante de tempo para passar pelo cruzamento. G1 e G2 são as distâncias do veículo à interseção. Quando dois veículos podem cruzar ao mesmo tempo, a prioridade é dada ao veículo em R1 com probabilidade P, e ao veículo em R2 com probabilidade 1 – P. Próximo à interseção, o veículo que tem prioridade se move com sua velocidade normal e o que não tem prioridade desacelera, como descrito no Algoritmo 1, onde d(i) = G(i) é a distância até a interseção.

4- Resultados e Discussão

Nesse relatório apresentamos somente alguns resultados. O paper enviado (que segue junto com presente relatório) apresenta outros resultados encontrados nesta pesquisa. Em todos eles consideramos que o comprimento total do circuito é 30km, e são obtidos depois de 20.000 instantes de tempo onde os 17.000 primeiros instantes de tempo são descartados, como usual. Para cada densidade, são distribuídos veículos aleatoriamente pela pista, com velocidade inicial *v(i,0)=0*. A densidade ρ é a porcentagem de células ocupadas na pista. Como no modelo NaSch tradicional, um veículo ocupa 7.5m, que dividimos em n células, onde o tamanho das células é dado por $l_c = (7.5/n)\ m$ e $A_{max} = n\ cell/s^2$. Para os resultados apresentados consideramos n = 5 e chamamos p_b de probabilidade de frenagem e P de probabilidade de prioridade do veículo em R1. Para essa discretização, definimos as seguintes políticas de aceleração/desaceleração:

(1) Agressiva: *β(10,30)* – Acel. média/s = $4\ cell/s^2 = 6\ m/s^2$
(2) Intermediária I: *β(20,28)* – Acel. média/s = $3\ cell/s^2 = 4.5\ m/s^2$
(3) Intermediária II: *β(28,20)* – Acel. média/s = $2\ cell/s^2 = 3\ m/s^2$
(4) Cautelosa: *β(30,10)* – Acel. média/s = $1\ cell/s^2 = 1.5\ m/s^2$

4.1 Comparando os perfis

A Figura 2 representa uma comparação entre os quatro diferentes perfis definidos e o modelo NaSch tradicional, com P = 1 e p_b = 0.01. Podemos observar a existência das quatro fases definidas por Marzoug [2]. Também vemos que a segunda fase (região platô) diminui com a aceleração média de cada perfil, mais pronunciadamente no perfil Cauteloso, mostrando a

influência de diferentes políticas de aceleração/desaceleração, considerando a mesma velocidade máxima da via, que nesse exemplo é 25cell/s = 135km/h.

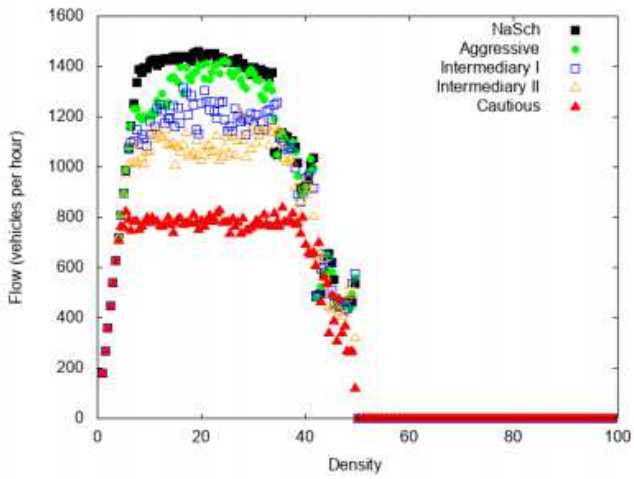


Figura 2 - Fluxo-Densidade: Comparação dos diferentes perfis

4.2 Resultados para diferentes prioridades na interseção

A Figura 3 mostra o diagrama Fluxo-Densidade para o perfil Agressivo, com p_b = 0.01 e $v_{max} = 25\ cell/s$, para diferentes valores da prioridade P. Podemos ver que o resultado reflete a descontinuidade entre a terceira e quarta fase descritas por Marzoug na densidade ρ = 50%, exceto quando P = 0.5, onde a descontinuidade ocorre aproximadamente quando ρ = 60%.

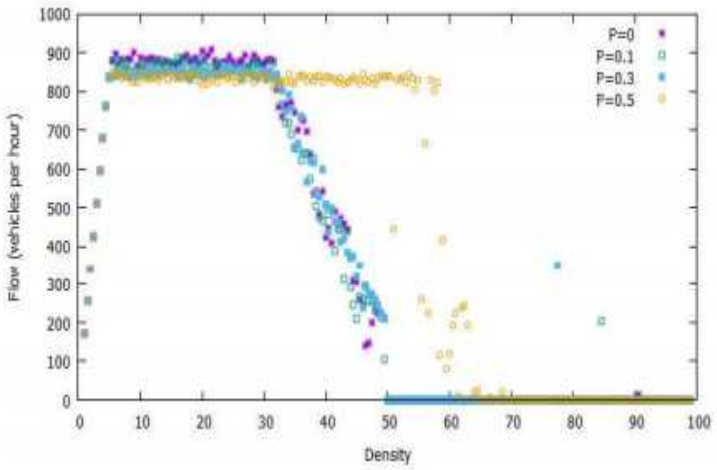


Figura 3 - Fluxo-Densidade: Diferentes P´s no perfil Agressivo

5- Conclusões e próximos passos

Durante o desenvolvimento dessa pesquisa temos mostrado que a forma como o motorista acelera e desacelera, que normalmente é um comportamento não observável em medições usuais e desconsiderado em modelos matemáticos, influencia de forma substancial no comportamento do tráfego, alterando sua dinâmica. Embora o modelo de cruzamento proposto por Marzoug [2] seja um modelo simplificado, foi possível avaliar nesse trabalho, como esses

diferentes perfis de motoristas podem afetar o fluxo de tráfego numa pista com interseção não sinalizada.

O modelo NaSch é um modelo explícito onde o motorista considera o movimento do veículo à sua frente apenas no instante de tempo anterior. Modelos mais realistas tentam representar, de alguma forma, esse movimento. São os chamados modelos de antecipação. Atualmente estamos trabalhando em uma proposta de um modelo que considere a antecipação, incluindo os diferentes comportamentos dos motoristas. Para isso estamos avaliando, na bibliografia, os principais modelos que apresentam essas características, como o proposto por Knospe [9], chamado de luz de freio. Pretendemos com isso, apresentar um modelo de antecipação que consiga considerar a antecipação e comportamentos de motoristas. O modelo proposto por Zamith [4,8] já considera antecipação e comportamento dos motoristas, mas precisa ser melhorado na parte da modelagem de frenagem e distância de segurança. Para isso, estamos introduzindo conceitos apresentados no modelo de luz de freio [9], que é um dos modelos de antecipação mais utilizado e citado na literatura. Com o conhecimento adquirido no tratamento desses diferentes perfis de aceleração/desaceleração, acreditamos que agora estamos realmente aptos a otimizar o modelo de antecipação já proposto pelo grupo de pesquisa.

Bibliografia

1. Nagel, K., Schreckenberg, M.: A cellular automaton model for freeway traffic. Journal de physique I 2(12), 2221-2229 (1992).
2. Marzoug,R., Ez-Zahraouy, H., Benyoussef, A.: Cellular automata traffic flow behavior at the intersection of two roads, Physica Scripta, 89(6), 1-7 (2014).
3. Zamith, M.P., Um modelo de autômato celular aplicado ao tráfego viário com múltiplos perfis de condutores, 2013, Tese de Doutorado, Pós-graduação em Computação, Instituto de Computação, UFF.
4. Zamith, M, Leal-Toledo, R. C. P., Clua, E., Toledo, E. M., Magalhes, G. V.: A new stochastic cellular automata model for traffic flow simulation with drivers behavior prediction. Journal of computational science 9, pp. 51-56 (2015).
5. I. M. Almeida , R. C. P. Leal-Toledo, E. M. Toledo, D. C. Cacau, e G. V. P. Magalhães, Drivers' behavior effects in the occurrence of dangerous situations which may lead to accidents, 5th Workshop on Traffic and Cellular Automata, LNCS 11115, 441-450, 2018.
6. Leal-Toledo**,** R. C. P., Magalhães, G.V. P., Miranda, I., Toledo, E. M., Traffic Flow Simulation Under the Influence of Different Acceleration Policies, Int'l Conf. Scientific Computing (CSC'17), pp 90-94 (2017).
7. Qi-Lang, L., Jiang, R.; Min, J.; Xie, J.-R.; Wang, B. H.: Phase diagrams of heterogeneous traffic flow at a single intersection in a deterministic Fukui-Ishibashi cellular automata traffic model. Europhysics Letters, 108(2), 28001-28008 (2014).
8. Nagatani, T.: Traffic states and fundamental diagram in cellular automaton model of vehicular traffic controlled by signals, Phy s.A: Statistical, Mechanics and its Applications,388(8), 1673-1681 (2009)
9. Knospe, W., Santen, L., Schadschneider, A., Schreckenberg, M.: Towards a realistic microscopic description of highway traffic. Journal of Physics A: Mathematical and general 33(48), pp. L477 (2000).

Plano de trabalho

1 Dados gerais

Título: Computação em máquinas não-confiáveis
Bolsista: Ricardo Luiz Cerqueira Júnior
Orientador: Fábio Borges de Oliveira
Bolsa: PIBIC
Vigência: 05/2019 - 07/2019

2 Objetivos

Objetivo Geral

1. Descrever técnicas usadas para proteção;
2. Estudar criptografia homomórfica;
3. Estudar DC-Nets;
4. Realizar análise e comparação entre criptografia homomórfica e DC-Nets.

Objetivos Específicos

1. Descrever algoritmos de criptografia homomórfica;
2. Descrever algoritmos de DC-Nets;
3. Estudar Teoria de Grupos;
4. Verificar o entendimento dos algoritmos através de implementação em Fortran.

3 Introdução

Com o avanço das tecnologias de comunicação e informação, a segurança e privacidade passam a ser controladas pelo mundo cibernético. Vulnerabilidades em tais tecnologias representam eminentes ameaças à segurança e privacidade, tanto no nível individual quanto coletivo. Pessoas podem ser individualmente lesadas e manipuladas. Um país pode ter sua infraestrutura crítica danificada

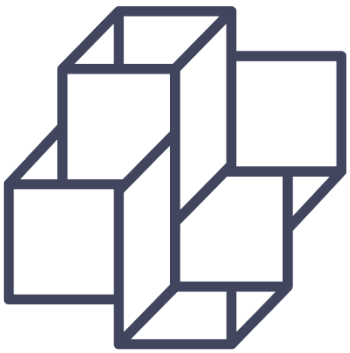
e uma eleição controlada. Diversas técnicas criptográficas são usadas para proteger a segurança e privacidade. Em particular, criptografia homomórfica e DC-Nets são técnicas usadas para garantir segurança e privacidade em diversos cenários de aplicação. Em especial, estamos interessados em um cenário onde usuários submetem seus dados sensíveis e programas secretos para rodarem em um supercomputador controlado por pessoas que podem ter interesses nos dados e programas.

4 Metodologia

1. Estudo Dirigido;
2. Apresentação de Seminário.

5 Referências bibliográficas

- William Stallings, Cryptography and Network Security, 2014, Prentice Hall
- Routo Terada, Segurança de Dados: Criptografia em Redes de Computadores, 2000, Edgard Blucher
- A Acar, H Aksu, AS Uluagac, M Conti, A survey on homomorphic encryption schemes: Theory and implementation, ACM Computing Surveys (CSUR), 2018
- David Chaum, The Dining Cryptographers Problem: Unconditional Sender and Recipient Untraceability, 1988, Journal of Cryptology
- S. Shokranian, M. Soares, H. Godinho, Teoria dos números, 1999, Ed. UNB
- F. Borges, Privacy-preserving data aggregation in smart metering systems, 'Energy: From Smart Metering to the Smart Grid', Chap. 2, 2016, IET



Laboratório Nacional de Computação Científica

Relatório de Atividades

MODELAGEM DE SISTEMAS T´ERMICOS COMPLEXOS

Thiago da Rocha Canella | Petrópolis, 19 de Julho de 2019

Objetivos:

Implementação eficaz de controle de temperatura e umidade de um ambiente controlado através do método de lógica Fuzzy, comparando os resultados obtidos com demais métodos, como exemplo o PID.

Introdução:

Baseando-se em um sistema de controle eletrônico/robótico, a pesquisa foi iniciada com base em um ambiente fechado controlando a temperatura através do controle da intensidade de uma lâmpada incandescente para atingir uma temperatura alvo. Posteriormente o protótipo de controle desta temperatura ganhou evoluções como controle de intensidade afinada, ventilação e portinhola para entrada de ar externo, visando ter controle rápido e eficaz.

Materiais e Métodos:

Como material de desenvolvimento foram usados:

- Arduino modelo Duamillenove
- Modulo sensor de temperatura DS18B20
- Cooler PWM Intel modelo E97379-001
- Motor de passo MG90S
- Módulo Triac (Fabricação própria)
- Módulo Relé de 2 estágios (Fabricação própria)
- Lâmpada incandescente 40W 110V
- Caixa de papelão 29cm X 17cm X 32cm

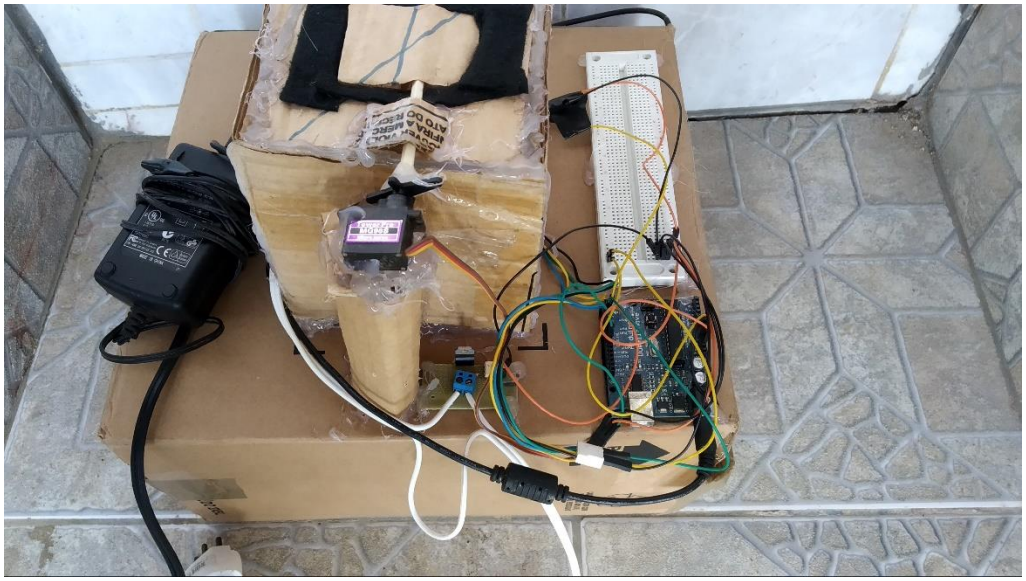


imagem do ultimo protótipo até a data deste relatório

Quanto à linguagem e metodologia de programação utilizada foi escolhido Python 3.7 rodando em um sistema operacional à escolha, podendo ser: Linux, Windows ou MacOs. Aplicação Python está responsável pela lógica de funcionamento e tratamento de dados, telemetria e registro de dados para análise.

Por sua vez, os dados assimilados pelos sensores e os acionamentos dos seus atuadores é responsabilidade do controle do Arduino, que é programado na sua linguagem Java de framework de sua IDE própria 'Arduino IDE'.

A comunicação entre os 2 é feita através de mensagens de texto, modulados em uma comunicação serial TTL pela sua comunicação USB com o computador.

Os softwares usados foram:

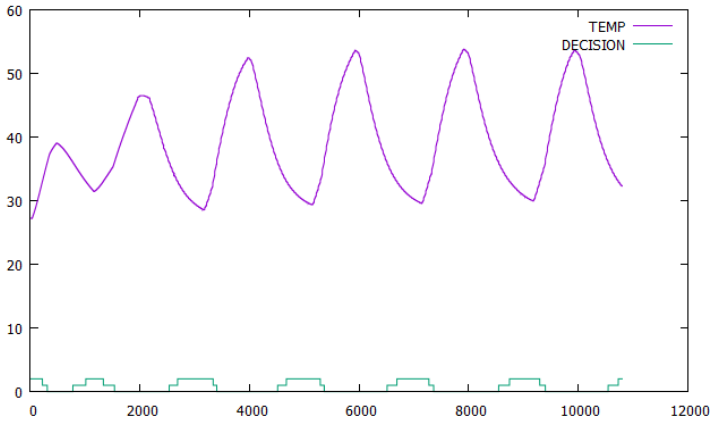
- Visual Studio Code
- Python 3.7.4
- Arduino IDE.

Resultados e Discussão:

Inicialmente o objetivo no primeiro protótipo foi a tentativa de controle da temperatura no interior da caixa com somente a lâmpada. Sendo usado o módulo de Relé para criar um ambiente de 2 estágios de intensidade, com o primeiro estágio com um diodo que permitia apenas passagem de metade da fase de 110vAC para metade de sua intensidade, e seu segundo estágio como uma ligação direta 110vAC para total intensidade da lâmpada.

Nos experimentos iniciais usou-se um método proporcional, tentando regular a temperatura através da irradiação do calor gerado no gás do interior da lâmpada através desses 2 estágios de intensidade para estabelecer a temperatura alvo almejada do código lógico.

Sabendo que haveria problemas nessa fase inicial, observou-se que a inércia do gás da lâmpada em conjunto com a dificuldade de dissipação da caixa fechada, produzindo assim uma oscilação muito maior do que o desejado de forma inesperada como mostrado no gráfico a seguir onde o objetivo era o alvo de 36º:



Com todo esse problema observado, ficou decidido a necessidade imediata da evolução do protótipo para um modelo onde poderia ter um controle fino da intensidade da lâmpada. Mesmo assim, a inércia do gás da lâmpada dificultou o alvo, permanecendo o problema relativo aos testes no antigo protótipo.

Logo em seguida foi implementado um sistema de resfriamento através do cooler e porta de abertura citados anteriormente. Tornando o sistema mais sofisticado e podendo aceitar comandos mais finos e precisos. Para qual foi adotado uma evolução no código que deixaria de ser proporcional e se tornaria um controle PID, com intuito de controle refinado individualmente de cada atuador do protótipo.

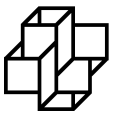
No momento está sendo feito um ajuste mais refinado no sistema PID, para que o funcionamento do protótipo seja validado, para que então seja colocado o modelo de lógica Fuzzy inicial em prática. Portanto sem validar o protótipo com um método mais simples, configuraria uma complexidade com excesso de incógnitas e variáveis que são extintos com a validação do modelo.

Conclusão:

Ainda em meio a testes e adequações, estamos evoluindo com a experiência, porém ainda não foi possível alcançar uma conclusão. Por hora está sendo estudado fuzzy mais a fundo e realizando mais testes

Bibliografia:

- Fuzzy Control (1998 - Kevin M. Passino, Stephen Yurkovich)
- Sistema de Controle de Temperatura para uma Estufa com Monitoramento via Aplicativo (2016 - Orlem L. D. Santos, Jo D. S. M. Júnior, Mendelsson R. M. Neves)
- https://www.circuitar.com.br/projetos/forno-pid-edison/
- Understanding and Design of an Arduino-based PID Controller (2016 - Dinesh Bista)
- https://www.researchgate.net/publication/260232896_Temperature_Control_using_Fuzzy_Logic



Laboratório
Nacional de
Computação
Científica

Projeto de Pesquisa de Iniciação Científica

Programa Institucional de Bolsas de Iniciação Científica - PIBIC

Inversão de Dados Sísmicos

Bolsista: Vinícius Theobaldo Jorge
Orientador: Marcio Rentes Borges

Petrópolis-RJ
19 de julho de 2019

Sumário

1 Identificação 1
2 Introdução 1
3 Metodologia 1
3.1 Método de Diferenças Finitas 2
4 Resultados Numéricos 2
4.1 Problema 1D 2
4.2 Fonte Sísmica 5
5 Conclusões 7

1 Identificação

- **Orientador:** Marcio Rentes Borges
  - Pesquisador Adjunto do Laboratório Nacional de Computação Científica (LNCC).
  - **E-mail:** *mrborges@lncc.br*
- **Bolsista:** Vinícius Theobaldo Jorge
  - Aluno do sétimo período do curso de graduação de Geologia da UFRJ.
  - **E-mail:** *vinicius.tj@gmail.com*

2 Introdução

A geofísica desempenha papel crucial na descoberta de novos reservatórios de petróleo e, ainda hoje, o principal uso de dados sísmicos é identificar a geometria dos refletores. Isso é possível porque as ondas sísmicas refletem nas interfaces entre materiais de diferentes propriedades acústicas (Frazer et al., 2008). Entretanto, recentemente, estudos sísmicos tem sido usados para a caracterização das rochas e monitoramento dos reservatórios (sísmica 4D), pela transformação de dados de reflexão em propriedades de rochas. Portanto, a inversão de dados sísmicos é uma ferramenta essencial na determinação das propriedades elásticas do subsolo.

O objetivo principal deste projeto é a inversão de dados sísmicos utilizando uma abordagem estocástica (Sen e Biswas, 2017). Entretanto, em sua primeira fase, o objetivo foi o estudo de métodos numéricos para aproximação da equação da onda. Este relatório apresenta a conclusão desta fase do projeto que culminou na construção de um código computacional em 3D para aproximar solução da equação da onda em meios elásticos heterogêneos.

3 Metodologia

O modelo físico considerado é um meio elástico heterogêneo pelo qual o som é propagado através de pequenas vibrações elásticas dado pela seguinte equação:

$$\frac{\partial^2}{\partial t^2} p\left(\vec{x},t\right) = \mathsf{K}\nabla\cdot\left[\frac{1}{\rho\left(\vec{x}\right)}\nabla p\left(\vec{x},t\right)\right] + \varphi\left(\vec{x},t\right), \tag{1}$$

onde p é a pressão, ρ a massa específica, K o módulo elástico e φ o termo de fonte.

Foi implementado e testado um código computacional, escrito em FORTRAN 90, para aproximar a solução da equação (1) utilizando o método de diferenças finitas de segunda ordem, em um domínio tridimensional.

3.1 Método de Diferenças Finitas

O método de diferenças finitas foi utilizado para aproximar a solução da equação (1). Os domínios, espacial e temporal, foram discretizados por um conjunto de pontos, nos quais a aproximação foi calculada. Mais especificamente, utilizamos uma malha espacial regular com espaçamentos Δx, Δy e Δz. Para as derivadas temporal e espacial foram utilizadas aproximações centradas de segunda ordem. Definindo $\mathsf{P}^n_{i,j,k}$ como uma aproximação para $p(x_i, y_j, z_k, t^n)$, onde os índices i, j, k se referem aos pontos da malha nas direções x, y, z, respectivamente, enquanto índice n representa o número do passo de tempo da simulação. Portanto, podemos escrever a seguinte aproximação:

$$\begin{aligned}
\mathsf{P}^{n+1}_{i,j,k} &= C\left(\frac{\Delta t}{\Delta x}\right)^2\left[\mathsf{P}^n_{i+1,j,k} - 2\mathsf{P}^n_{i,j,k} + \mathsf{P}^n_{i-1,j,k}\right] + \\
&\quad C\left(\frac{\Delta t}{\Delta y}\right)^2\left[\mathsf{P}^n_{i,j+1,k} - 2\mathsf{P}^n_{i,j,k} + \mathsf{P}^n_{i,j-1,k}\right] + \\
&\quad C\left(\frac{\Delta t}{\Delta z}\right)^2\left[\mathsf{P}^n_{i,j,k+1} - 2\mathsf{P}^n_{i,j,k} + \mathsf{P}^n_{i,j,k-1}\right] + \\
&\quad 2\mathsf{P}^n_{i,j,k} - \mathsf{P}^{n-1}_{i,j,k} - \Delta t^2\delta(i,i_\varphi)\delta(j,j_\varphi)\delta(k,k_\varphi)\varphi^n
\end{aligned} \tag{2}$$

onde $C = c^2_{i,j,k} = \left(\dfrac{\mathsf{K}(x_i, y_j, z_k)}{\rho(x_i, y_j, z_k)}\right)^2$. Os pontos i_φ, j_φ, k_φ indicam a posição da fonte nos nós da malha, portanto, $\delta(\alpha, \alpha_\varphi) = 1$ caso $\alpha = \alpha_\varphi$ e $\delta(\alpha, \alpha_\varphi) = 0$, caso contrário, onde $\alpha = i, j, k$.

4 Resultados Numéricos

Nesta seção, apresentamos os resultados obtidos para 2 experimentos que serão descritos a seguir.

4.1 Problema 1D

Considere um material elástico, com velocidade de propagação $c = \sqrt{\dfrac{\mathsf{K}}{\rho}}$, no domínio unidimensional $[L_0,\ L_1]$, com base na equação (1), o seguinte problema matemático foi

utilizado para os testes:

$$p_{tt} = c^2 p_{xx} + \varphi \tag{3}$$

em conjunto com as seguintes condições iniciais e de contorno:

$$\begin{array}{rcll} p(x,0) & = & f(x), & x \in [L_0,\ L_1] \\ p_x(x,0) & = & 0, & x \in [L_0,\ L_1] \\ p(L_1,t) & = & 0, & t \in (0,\ T] \\ p(L_2,0) & = & 0, & t \in (0,\ T] \end{array} . \tag{4}$$

A solução da Eq. (3) associada com condições dadas na Eq. (4), foi aproximada utilizando o método de diferenças finitas centradas. Desconsiderando o termo de fonte e utilizando aproximações de segunda ordem no tempo e segunda e quarta ordens no espaço, obtemos os seguintes métodos:

- Aproximação de segunda ordem:

$$P_i^{n+1} = -P_i^{n-1} + 2P_i^n + C_i^2 \left(P_{i+1}^n - 2P_i^n + P_{i-1}^n\right) \tag{5}$$

- Aproximação de quarta ordem:

$$\begin{aligned} P_i^{n+1} = & -P_i^{n-1} + 2P_i^n + \frac{C_i^2}{12}\left[-P_{i-2}^n + 16P_{i-1}^n - \right. \\ & \left. 30P_i^n + 16P_{i+1}^n - P_{i+2}^n\right] \end{aligned} \tag{6}$$

onde $C = c\Delta t/\Delta x$, $L_0 = -40$, $L_1 = 40$ e $f(x) = \exp(-x^2)$. Para visualizar a reflexão e transmissão das ondas, utilizamos dois valores de velocidade de propagação:

$$c(x) = \begin{cases} 2, & L_0 \leqslant x \leqslant 20, \\ 1, & 20 < x \leqslant L_1. \end{cases} \tag{7}$$

O domínio foi discretizado em uma malha de 501 nós e, para garantir estabilidade do método, consideramos $\Delta t = \dfrac{1}{2}\dfrac{\Delta x}{\max\{c\}}$.

Quando uma onda propagante se depara com uma barreira, ou uma interface entre dois meios diferentes (mudança de velocidades dada em (7)), podem ocorrer dois

fenômenos: reflexão e transmissão. As relações entre a onda incidente (Ai) e as ondas transmitida (At) e refletida (Ar) são dadas por:

$$\frac{\mathsf{Ar}}{\mathsf{Ai}} = R \qquad \text{e} \qquad \frac{\mathsf{At}}{\mathsf{Ai}} = T,$$

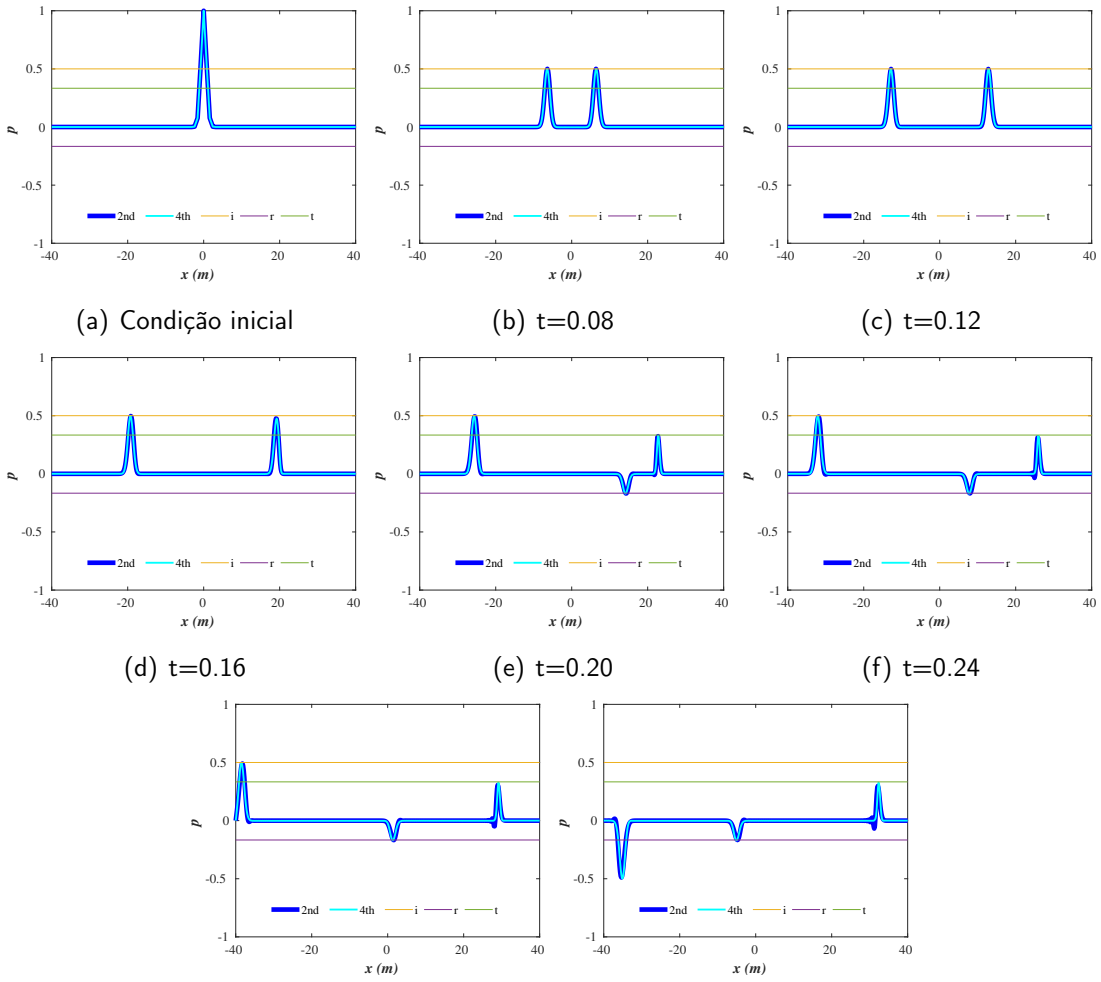
onde R e T são os coeficientes de reflexão e transmissão, respectivamente, determinados por:

$$\frac{\mathsf{Ar}}{\mathsf{Ai}} = R = \frac{v_2 - v_1}{v_1 + v_2}$$

e

$$\frac{\mathsf{At}}{\mathsf{Ai}} = T = \frac{2v_2}{v_1 + v_2}.$$

Essas relações foram usadas para prever o comportamento das ondas e são representadas na Figura 1. Podemos ver que as soluções aproximadas das ondas comportaram-se segundo as previsões teóricas.



(a) Condição inicial (b) t=0.08 (c) t=0.12

(d) t=0.16 (e) t=0.20 (f) t=0.24

(g) t=0.28 (h) t=0.32

Figura 1: Solução aproximada da pressão.

4.2 Fonte Sísmica

Neste experimento consideramos um domínio de $3000 \times 3000 \times 3000m^3$ no qual uma fonte sísmica artificial foi colocada no centro deste. O material apresenta duas camadas com propriedades distintas na direção z. De baixo para cima, nos primeiros $1000m$ a velocidade de propagação da onda é de $2200m/s$, nos $2000m$ restantes a velocidade de propagação é de $2800m/s$ (Figura 3).

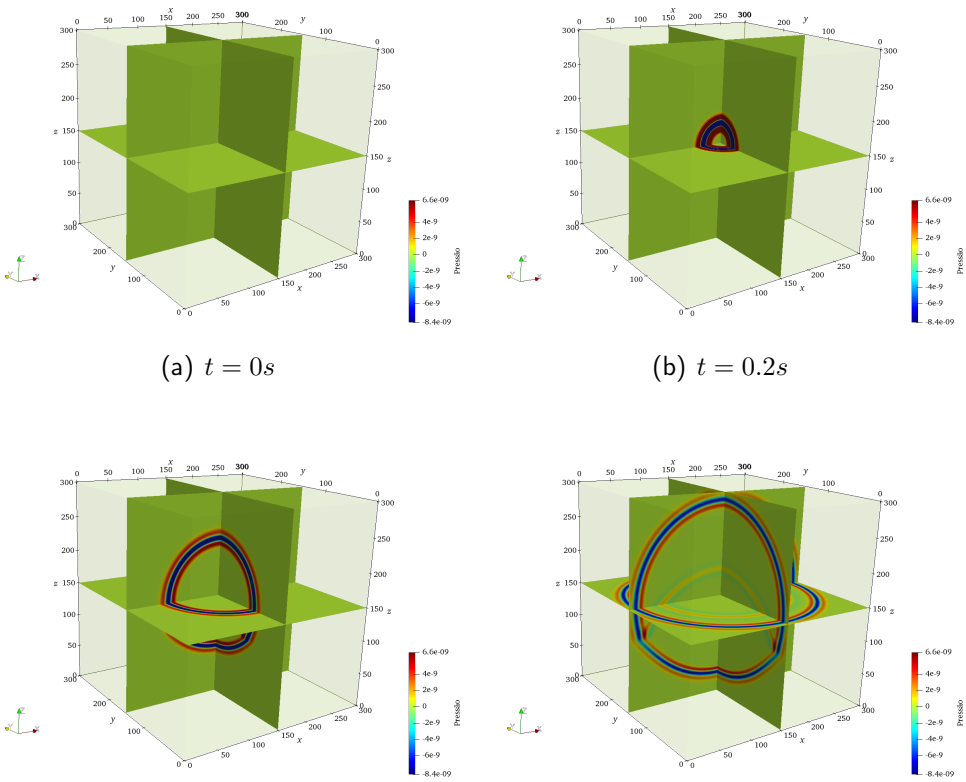
O modelo matemático utilizado para a fonte sísmica foi obtido a partir da segunda derivada da função Gaussiana, conforme descrito em (CUNHA, 1997) e é dado por:

$$\varphi(t) = \left[1 - 2\pi(\pi f_c t_d)^2\right] e^{-\pi(\pi f_c t_d)}, \tag{8}$$

onde $t_d = t - \dfrac{2\sqrt{\pi}}{f_{cut}}$ é o tempo defasado, necessário para deslocar a função para a direita, garantindo que $\varphi(t) = 0 \quad \forall t < 0$, e f_c é a frequência central da fonte, dada por:

$$f_c = \frac{f_{cut}}{3\sqrt{\pi}}. \tag{9}$$

A Figura 2 mostra a propagação da onda sísmica em quatro tempos distintos. Podemos observar que a onda, ao atingir a camada com propriedade diferente, parte dela é refletida e parte é transmitida com velocidade diferente (Figura 2(d)). Uma visualização mais detalhada desse resultado é dada na Figura 3.



(a) $t = 0s$ (b) $t = 0.2s$

(c) $t = 0.4s$ (d) $t = 0.6s$

Figura 2: Pressão em quatro tempos distintos. Os eixos x, y e z indicam o número de pontos da malha.

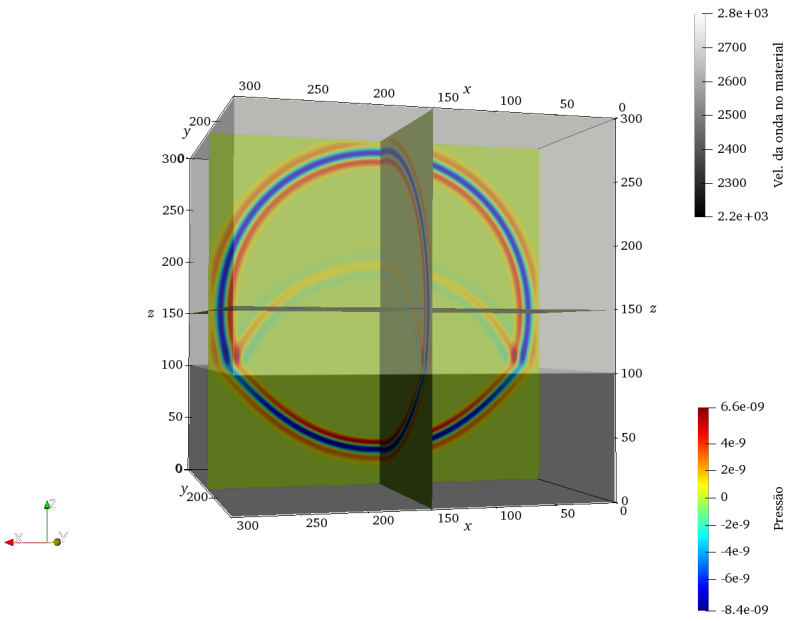


Figura 3: Transmissão e reflexão da onda sísmica ao atravessar diferentes materiais.

5 Conclusões

O método de diferenças finitas proporcionou uma boa aproximação da solução da Eq. (3), capturando o comportamento físico esperado para ambos problemas estudados. Com isso obtivemos um programa que está qualificado para usar dados reais de estudos relacionados a toda cadeia produtiva do petróleo.

Referências

P. E. M. CUNHA. *Estratégias Eficientes para Migração Reversa no Tempo Pré- empilhamento 3-D em Profundidade pelo Método das Diferenças Finitas*. Tese de Doutorado, P.P.P.G., Universidade Federal da Bahia, 1997.

B. Frazer, A. Bruun, K. B. Rasmussen, J. C. Alfaro, A. Cooke, D. Cooke, D. Salter, R. Godfrey, D. Lowden, S. McHugo, H. Özdemir, S. Pickering, F. G. Anderson Pineda, J. Herwanger, S. Volterrani, A. Murneddu, A. Rasmussen, e R. Roberts. Seismic inversion: Reading between the lines. *Oilfield Rev*, 20(1):42–63, 2008.

M. K. Sen e R. Biswas. Transdimensional seismic inversion using the reversible jump hamiltonian monte carlo algorithm. *Geophysics*, 82(3):119–134, 2017.

Relatório de atividades

Dados gerais

Verificação de propriedades físicas em Lentes Gravitacionais em Big Data

Bolsista: Viviane de Mattos Matioli

Orientador: Fabio Porto

Bolsa de iniciação científica (PIBIC) - 1/4/2019 a 31/7/2019

Objetivos

O objetivo do projeto é dar continuidade e auxiliar na pesquisa da área de astronomia computacional desenvolvida pelo grupo, relacionada mais especificamente à detecção de lentes gravitacionais a partir da análise de catálogos astronômicos. Como estudante de astronomia, devo ajudar na definição e implementação de critérios físicos que permitam a identificação de ocorrências deste fenômeno.

1. Introdução

De acordo com a Teoria da Relatividade Geral de Einstein, o efeito observado da gravidade é na verdade causado pela curvatura do espaço-tempo devido a concentrações de massa/energia. Grandes concentrações de massa (como galáxias e aglomerados) fazem com que o espaço-tempo ao seu redor se curve, o que afeta a trajetória de corpos próximos e também da luz.

Assim, quando a luz proveniente de um objeto astronômico (fonte) passa próxima a um outro corpo massivo (lente) antes de chegar até os observadores, ela será desviada. O efeito observado por nós são distorções na imagem da fonte (lenteamento fraco), ou até formação de várias imagens da mesma fonte (lenteamento forte).

Estas várias imagens são observadas e aparecem como diferentes entradas nos catálogos, e é necessária uma análise de suas características para determinar se elas na verdade representam o mesmo objeto. Algumas propriedades que podem ser utilizadas são coordenadas, redshift

(deslocamento do radiação emitida para o vermelho devido ao afastamento do objeto causado pela expansão do universo), fluxo nos diferentes filtros e espectro ([1] e [2]).

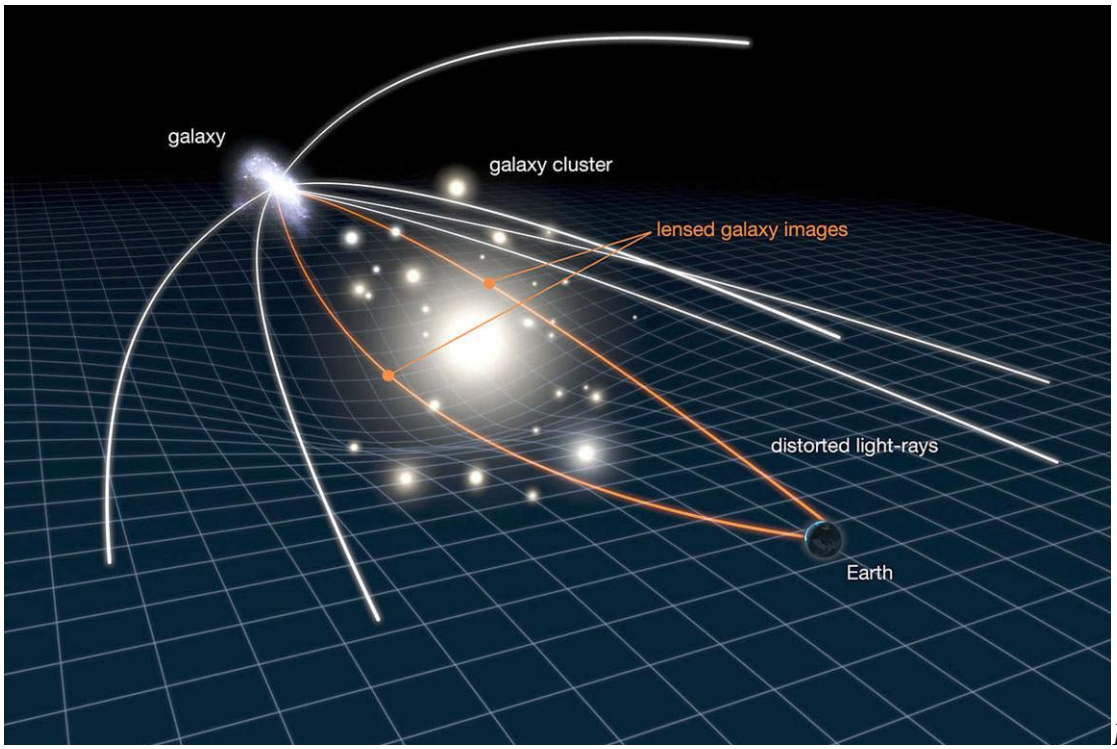


Figura 1: Esquema de uma lente gravitacional em que a luz de uma galáxia fonte é desviada por um aglomerado de galáxias

Até hoje há um número relativamente baixo de sistemas de lentes gravitacionais conhecidas quando comparado à quantidade crescente de dados astronômicos disponíveis. Assim os catálogos já existentes podem conter mais exemplares escondidos, e o objetivo do projeto é identificá-los.

Esse tipo de lente gravitacional é de grande importância na astronomia pois possui várias aplicações. Como o desvio da luz depende apenas da concentração de massa do objeto massivo causador da lente, o estudo de como este desvio ocorre é usado para determinar a distribuição de massa de objetos agindo como lentes. Um outro tópico ainda em aberto na física é a matéria escura, como este tipo de matéria só interage gravitacionalmente e não pode ser observado, lentes gravitacionais representam uma boa forma de estudar seus efeitos. Além disso, as lentes podem ser usadas também como mais uma maneira de estimar parâmetros cosmológicos como a constante de Hubble, relacionada à expansão acelerada do universo, que ainda possui discrepância entre valores encontrados com diferentes métodos, e por isso é um assunto importante atualmente.

2. Material e métodos

Neste início do projeto, o maior foco foi em estudar e entender as ferramentas (Python, Spark) e a teoria das lentes, para definir como implementar outros critérios ao projeto já existente de identificação geométrica para encontrar lentes – Constellation Queries [3], e outros métodos que poderiam ser utilizados para esta busca. Utilizando principalmente dados do Sloan Digital Sky Server.

Como os candidatos a lentes devem ter o mesmo redshift, considerando incertezas da medição, e estarem próximos espacialmente, foi estudado e implementado em python um método utilizando o algoritmo DBSCAN (Density-based spatial clustering of applications with noise) para separar o dataset em grupos com o mesmo redshift, considerando um parâmetro de tolerância, e em cada grupo foi aplicado novamente o DBSCAN para separá-los em grupos com componentes de distância de até 5 arcosegundos entre si.

Auxiliando no projeto das *Constellation Queries*, fiz uma análise de uma partição do dataset que tomava muito mais tempo na execução do programa comparada às suas vizinhas, para tentar encontrar a causa do problema e verificar se estava relacionada a alguma característica astrofísica particular daquela partição.

Atualmente estou estudando a relação esperada entre redshift da fonte e da lente e a diferença destes valores e a separação angular das múltiplas imagens formadas, a partir da teoria, da análise de dados de lentes já identificadas do survey CASTLeS (https://cfa.harvard.edu/castles), e de informações de simulações de lentes gravitacionais como o projeto *Bologna Lens Factory* (http://metcalf1.difa.unibo.it/blf-portal/index.html).

3. Resultados e Discussão

Com o algoritmo do DBSCAN obtivemos como resultado alguns grupos muito mais densos que os outros, devido a distribuição de redshift dos objetos da base de dados, que estão muito concentrados próximos de 0. Este algoritmo possui dois parâmetros, tolerância para redshift e para distância espacial, que podem ser melhor ajustados aplicando o programa em dados conhecidos como as lentes de CASTLeS.

Após a separação do dataset nos grupos, pode ser feita a combinação de objetos 4 a 4, para procurar por cruzes de Einstein (quando há a formação de 4 imagens com a lente no meio, formando uma cruz), por exemplo, porém este é um procedimento demorado e pode ser estudada a sua execução em paralelo utilizando Spark.

Em relação ao *Constellation Queries* e a análise da partição mais demorada, não foi identificada nenhuma diferença significativa entre ela e suas vizinhas na distribuição espacial e de redshift de seus objetos. O problema pode estar relacionado à utilização do critério de cores iguais para as diferentes imagens produzidas pela lente, e está sendo estudado por outros membros do grupo que desenvolveram o algoritmo.

A implementação de mais critérios como um limite para a diferença de redshift de lente e fonte, e da distribuição espacial das imagens, pode reduzir o número de candidatos e ajudar na identificação de lentes pelos 2 métodos. Assim, estou estudando para definir quais seriam estas relações e como implementá-las nos algoritmos.

4. Conclusões

De acordo com o planejamento, o início do projeto foi dedicado principalmente a familiarização com o tema e ferramentas, estudando artigos e linguagens de programação necessárias para a sua realização. Com implementação de alguns códigos como tentativas iniciais de desenvolvimento de um outro método para identificar lentes utilizando clustering.

Nas próximas etapas devem ser definidos de forma mais concreta quais outros critérios podemos utilizar no projeto de *Constellation Queries* e como implementá-los e testá-lo.

5. Referências bibliográficas

[1] Joachim Wambsganss, “Gravitational Lensing in Astronomy”. Living Rev. Relativity, 1, 1998

[2] P. Schneider, C. Kochanek, J.Wambsganss. Gravitational Lensing: Strong,Weak andMicro, 2006

[3] Fabio Porto, Amir Khatibi, João N. Rittmeyer, Eduardo Ogasawara, Patrick Valduriez, Dennis E. Shasha, Constellation Queries in Big Data, SBBD 2018.

1. Dados gerais

TÍTULO: MÉTODOS NUMÉRICOS PARA O ESCOAMENTOS BIFÁSICO EM MEIOS POROSOS HETEROGÊNEOS EM AMBIENTES COMPUTACIONAIS DE ARQUITETURA DE MEMÓRIA HÍBRIDA.

Bolsista: Weber Guilherme Dias Ribeiro
Orientadora: Carla Osthoff Barros
Co-orientador: Frederico Luís Cabral
Programa Institucional de Bolsas de Iniciação Científica – PIBIC.
Esse relatório cobre as atividades realizadas no período de 1 de agosto de 2018 a 31 de julho de 2019.

2. Objetivos

O objetivo do presente trabalho é desenvolver conhecimento na área de processamento de alto desempenho para avaliar e otimizar uma aplicação de computação científica em máquinas de arquitetura paralela híbrida, com o auxílio das ferramentas de análise de código.

3. Introdução

A proposta desse projeto de iniciação científica é dar continuidade ao projeto de otimização do código "MÉTODOS NUMÉRICOS PARA O ESCOAMENTO BIFÁSICO EM MEIOS POROSOS HETEROGÊNEOS", desenvolvido pelo grupo de pesquisa em Óleo e Gás do LNCC, em ambientes computacionais de arquitetura composta por processadores multicores interligados em um ambiente de memória distribuída.

O grupo de pesquisa deste projeto decidiu implementar uma reestruturação do código legado, em um código moderno e estruturado em forma modular seguindo os conceitos de programação orientada objeto. O objetivo é desenvolver um código com tecnologia de software escalável para ser executado de forma eficiente e de fácil manutenção em um ambiente de supercomputação.

No atual momento, o trabalho de iniciação científica tem como objetivo desenvolver o conhecimento necessário para compreender a estrutura do código que está sendo reestruturado, mas especificamente em relação às técnicas e algoritmos necessários no desenvolvimento da solução de equações diferenciais ordinárias e parciais clássicas que são aplicáveis na modelagem de problemas de engenharia e ciências, utilizando métodos computacionais sequenciais e paralelos bem como solução de sistemas lineares e não lineares que são necessários à solução das equações em questão.

Na seção 5, Resultados e Discussão, serão apresentados os estudos informações extraídas sobre os temas estudados.

4. Material e Métodos ou Metodologia

Para essa etapa do projeto de pesquisa, o material e a metodologia ficaram restritos ao estudo de livros e artigos. Foram selecionados capítulos específicos dentro de livros didáticos comumente utilizados em disciplinas relacionadas à métodos numéricos e processamento em paralelo, bem como artigos que se relacionam parcial ou totalmente com o conteúdo que foi estudado nessa etapado projeto.

Foram realizadas sínteses acerca de todo o material estudado, com o objetivo de eficiência didática e consulta futuras.Todo material utilizado, livros e artigos, estão citados na seção 7, referências bibliográficas.

5. Resultados e Discussão

A seguir serão apresentados os resultados das pesquisas desenvolvidas para a realização das etapas previstas no plano de trabalho. As etapas de desenvolvimento apresentados na proposta de trabalho tiveram que ser alteradas porque o grupo de pesquisa decidiu reestruturar o código legado em um código moderno e

estruturado de forma modular segundo os conceitos de programação orientada objeto.

As alterações atingiram os planos de trabalho relativos às etapas 3 e 4 do plano de trabalho, que eram relacionada com os passos de otimização do código legado.

5.1) Primeira etapa do plano de trabalho:

Como passo inicial, da proposta de trabalho, foram realizados estudos a respeito dos conceitos, básicos e avançados, das arquiteturas Multicore e de processamento massivamente paralelo utilizando a arquiteturaIntel®Haswell™/Broadwell™/Skylake™ e Xeon Phi™.

Elementos base analisados:

- Histórico da arquitetura Intel
- Hyper-Threading
- Pipeline
-Superescalar
- Vetorização
- Paralelismo de Microarquitetura e de Processamento
- Aspectos, Características e Comparativos entre as microarquiteturas Core: Haswell™ / Broadwell™ / Skylake™
- Análise do Coprocessador Xeon Phi (KNC – Knight Corner™ e KNL - Knight Landing™)
- Histórico da arquitetura Intel® Xeon Scalable™, Extensões do Haswell / Broadwell / Skylake Xeon Phi™.

Algumas tecnologias de Paralelismo implementadas nos microprocessadores Intel®:

-Pipeline: é uma melhoria organizacional do processador. O Pipeline quebra o ciclo da instrução em um número de estágios separados que ocorrem em sequência como em uma linha de montagem. É Bem claro que esse processo acelera a execução da instrução, mas possui deficiências.

-Superscalar: Um processador superscalar é aquele em que múltiplos e independentes *Pipelines* de instruções são usados. Cada pipeline consiste de múltiplos estágios, de tal forma que cada *Pipeline* possa lidar com múltiplas instruções ao mesmo tempo. Os pipelines múltiplos introduzem um nível de paralelismo, possibilitando que múltiplos, fluxos de instruções sejam processados ao mesmo tempo. Um processador superscalar explora o que é conhecido como paralelismo em nível de instrução.

-Intel® AVX: O Advanced Vector Extension (AVX), são extensões da arquitetura do conjunto de instruções x86, propostas pela Intel em março de 2008 e primeiramente suportada pela Intel com o processador Sandy Bridge no início de 2011.Atualmente se encontra na versão AVX-512.

5.2) Segunda etapa do plano de trabalho:

A seguir será descrito o estudo desenvolvido na etapa 2 do projeto de trabalho, relacionado com os conceitos de programação paralela, mais especificamente, com o modelo de programação de memória compartilhada OpenMP e com o modelo de programação de memória distribuída MPI.

Computação Paralela:

Memória Compartilhada: OpenMultiProcessing (OpenMP) é uma tecnologia de programação paralela que permite dividir o trabalho em threads. Está disponível para diversas linguagens, incluindo a linguagem C e Fortran. Os ganhos podem ser de N vezes no tempo de execução (ou N vezes de speed-up) quando um programa paralelizado é executado em um processador de N cores.

Seu princípio de paralelismo é baseado em compartilhamento de memória, com execução simultânea dos trechos de código em forma de threads, estas podem ser executadas de forma independente nos núcleos dos processadores e podem compartilhar informações entre eles, armazenadas na memória.

O gerenciamento das threads obedece ao princípio fork-join, o qual consiste em um processo de inicialização (fork), execução, e finalização das mesmas (join). As threads são reunidas em grupos e são inicializadas. A execução de um próximo grupo só poderá ocorrer quando todas as threads do grupo anterior forem concluídas. O gerenciamento do grupo é feito a partir da thread master.

Diretivas de configuração:

Diretiva Parallel:Esta diretiva indica ao compilador que o trecho compreendido deve ser executado por todas as threads disponíveis, é criado um grupo com todas as threads e com o mesmo trecho de código. As threads dentro do mesmo grupo são inicializadas, e a master só permite a inicialização de outro grupo, com outra configuração, quando todas as threads desse grupo são concluídas, obedecendo a barreira de sincronização implícita descrita anteriormente. algumas cláusulas: if, private, shared, default, reduction, num_threads.

Diretiva FOR:A diretiva for é usada para configuração de trechos com loop do tipo for, pode-se trabalhar cada iteração do loop. Seguem as descrições de algumas cláusulas: schedule, podendo ser static, dynamic, guided, runtime, auto, nowait, ordered.

Diretiva Sections:Permite definir que trechos diferentes podem ser executados simultaneamente,define que o trecho seja serial.

Diretivas de sincronização: As diretivas de sincronização são úteis para determinar como certos trechos tenham garantia de exclusividade de execução.

Diretiva Master O trecho só pode ser executado pelo mestre do grupo, Diretiva Critical trecho pode ser replicado para todas as threads, Diretiva Barrier cria uma barreira de sincronização explícita.

Funções do OpenMP: possuem a finalidade de realizar configuração das threads durante a execução, omp_set_num_threads, intomp_get_num_threads, intomp_get_thread_num, intomp_in_parallel, omp_set_dynamic, intomp_get_dynamic.

Variáveis de ambiente: são inicializadas dentro do ambiente de execução, omp_schedule, omp_num_threads, omp_dynamic, omp_proc_bind.

Memória Distribuída:MessagePassing Interface (MPI) é uma tecnologia de programação paralela que permite dividir o trabalho em processos que são executados em diferentes computadores. Uma aplicação é constituída por um ou mais processos dependendo da implementação. Os processos se comunicam, através de funções de envio e recebimento de mensagens. Algo muito importante em programações paralelas é a comunicação de dados entre processos paralelos e o balanceamento da carga, que é uma forma de dizer quanto cada processo irá "trabalhar".

Alguns conceitos que definem elementos do ambiente de comunicação do MPI: Processo, é um programa distribuído, possui réplicas em cada um dos nós, Rank, é o identificador único de um processo, Grupo, conjunto de processos que podem trocar mensagens entre si, Comunicador, é um identificador de grupo, representa um domínio, Comunicação ponto-a-ponto, funções de comunicação entre dois processos, Comunicação coletiva, funções de comunicação de um para vários processos.

A respeito da sincronização dos dados transmitidos em comunicações ponto-a-ponto, o MPI possui suporte aos tipos: bloqueante ou não-bloqueante. A comunicação bloqueante tem por base apenas permitir a próxima instrução quando o receptor envia uma confirmação para o emissor. Caso a comunicação seja não-bloqueante, não há confirmação de recebimento, e a próxima instrução será executada independente do término da transferência. Caso a comunicação seja do tipo coletiva, as funções são todas não bloqueantes, cabendo ao usuário utilizar funções de barreira para forçar sincronização entre os processos. O protocolo suporta ainda diversos paradigmas de controle do fluxo como o mestre-escravo.

O modelo de programação consiste em configuração explicita do paralelismo, e defini como a transmissão de dados se comportará. As funções de envio e recebimento utilizam tipos próprios de dados para paralelismo de comunicação.

O MPI possui um conjunto de funções básicas: MPI_Init, MPI_Finalize, MPI_Comm_size, MPI_Comm_rank, MPI_Get_processor_name, MPI_Send, MPI_Recv, MPI_Bcast, MPI_Barrier, MPI_Scatter, MPI_Gather, MPI_Scatter.

Arquiteturas Híbridas

É a Implementação de programação paralela híbrida que utiliza memória distribuída com memória compartilhada, abordando os padrões de configuração para execução paralela de uma aplicação. Estas configurações são importantes para otimizar os trechos de software que serão integrados às arquiteturas de processamento durante o processo de desenvolvimento.

5.3) Terceira etapa do plano de trabalho:

Conforme justificado anteriormente, as etapas de desenvolvimento apresentados na proposta de trabalho tiveram que ser alteradas porque o grupo de pesquisa decidiu reestruturar o código legado em um código moderno e estruturado de forma modular segundo os conceitos de programação orientada a objeto.

As alterações foram realizadas para adquirir entendimento da estrutura dos métodos numéricos implementados no novo código. Como primeiro passo para o entendimento dos métodos numéricos e das suas respectivas equações diferencias, foi realizado um estudo dos mecanismos e leis que regem o funcionamento dos fenômenos físicos que ocorrem na natureza, realizado através de observações e formulação de definições e equações que tornam possíveis a criação de modelos matemáticos utilizados na solução e equacionamento de problemas, questões mais complexas envolvem o equacionamento através de diferenciações e a maioria deles com soluções analíticas extremamente complexas ou até mesmo inviáveis ou impossíveis, desta forma, a solução é obtida de maneira aproximada com o auxílio de métodos numéricos.

Os métodos numéricos buscam a solução de problema analíticos utilizando a discretização do domínio contínuo oriundo do equacionamento analítico, definindo condições de contorno e através de aproximações obtêm-se equações mais simplificadas, lineares ou não lineares, para a solução aproximada do problema, no atual estágio do estudo o método numérico analisado é o Método das Diferenças Finitas.

A) Equações Diferenciais Ordinárias: tradicionalmente equações diferenciais ordinárias de primeira ordem devem satisfazer a uma única condição inicial. Sendo (P.V.I – Problema do Valor Inicial) ou (P.V.L - Problema do Valor Limite). No estudo utilizaremos EDOs de segunda ordem, na forma:

$$y'' = f(x, y, y'), \qquad a \leq x \leq b$$

$$y(a) = \alpha \quad e \quad y(b) = \beta$$

Solução sequencial em MDF (Método das Diferenças Finitas) para EDOs: é um método de solução de equações diferenciais baseadas na aproximação de derivadas por diferenças finitas na qual a fórmula de aproximação obtém-se da Série de Taylor da função derivada. No presente estudo a aproximação numérica de derivadas será a diferença central.

$$y'(i) = \frac{y(i+1) - y(i-1)}{2h} \; y''; \; (i) = \frac{y(i+1) - 2y(i) + y(i-1)}{h^2}; y(a) = \alpha \quad e \quad y(b) = \beta$$

Algoritmo de solução sequencial em Diferenças Finitas para EDOs Lineares:

$$h = \frac{(b-a)}{(n+1)}$$

$$wo = \alpha \; (condição\ limite)$$

$$wn+1 = \beta \; (condição\ limite)$$

PARA i=1 até n

wi = 0

bi = -h2r(a + ih)

$$b1 = b1 + (1 + \frac{h}{2}\, p(a+h)\, wo$$

$$bn = bn + (1 + \frac{h}{2}\, p(b-h)\, wn+1$$

PARA j=1 até n-1 e i=j+1

$$Ai,j = -1 - \frac{h}{2} p(a+ih)$$

PARA j=1 até n-1 e i=j

$$Ai,j - 2 + h.h.q(a+ih)$$

PARA j=2 até n-1 e i=j-1

$$Ai,j = -1 + \frac{h}{2} p(a+ih)$$

Resolva o sistema linear Aw = b

Algoritmo de solução sequencial em Diferenças Finitas para EDOs não Lineares:

$$h = \frac{(b-a)}{(n+1)}$$

$$wo = \alpha \; (condição\ limite)$$

$$wn+1 = \beta \; (condição\ limite)$$

PARA i=1 até n

$$wi = \alpha + i.h.\frac{\beta - \alpha}{b-a}$$

Resolver F(w) = 0

Calcula os elementos de F(w1,...,wn) para a EDO não linear

PARA i=1 até n

$$Fi = -w(i-1) + 2w(i) + h.h.f(a+ih,$$

Calcula os elementos de J(w1,...,wn) para a EDO não linear

PARA j=1 até n-1 e i=j+1

$$Ji,j-1 - \frac{h}{2}.fy'(a+ih, wi, \frac{w(i+1) - w(i-1)}{2.h}$$

PARA j=1 até n e i=j

$$Ji,j = 2 + h.h.fy(a + ih, wi, \frac{w(i+1) - w(i-1)}{2.h}$$

PARA j=2 até n e i=j-1

$$Ji,j = -1 + \frac{h}{2}.fy'(a + ih, wi, \frac{w(i+1) - w(i-1)}{2.h}$$

B) Equações Diferenciais Parciais (EDPs): as equações diferenciais parciais podem ser Elípticas, Parabólicas ou Hiperbólicas, cada tipo se enquadra melhor para modelar determinado fenômeno físico, transiente ou estacionário.

Para fenômenos em estado de equilíbrio (estacionários) se enquadram as EDPs elípticas e para os fenômenos transientes se enquadram as EDPs parabólicas e hiperbólicas, sendo que as parabólicas consideram o fenômeno dissipativo (perda de energia) e as hiperbólicas consideram o fenômeno não dissipativo.

EDPs. Elípticas: condições não se alteram em função do tempo.

$$\nabla^2 u(x,y) \equiv \frac{\partial^2 u}{\partial x^2}(x,y) + \frac{\partial^2 u}{\partial y^2}(x,y) = f(x,y).$$

Figura 1-Poisson Cartesiana sem condições de contorno aplicadas

Algoritmo de solução sequencial em Diferenças Finitas para EDPs não Lineares:

$$h = \frac{(b-a)}{(n+1)}$$

$$k = \frac{(d-c)}{(m+1)}$$

PARA i=1 até n

PARA j=1 até m

$$w(i \quad 1),(m \mid j) = 0$$

$$b(i-1),(m+j) = -h.h.f(a+ih, c+jk)$$

PARA j=1 até m

$$bj = bj + g(a, c+jk)$$

$$b(n-1),(m+j) = b(n-1),(m+j) + g(b, c+jk)$$

λ = h.h / k.k

μ = 2(1 + λ)

PARA j=1 até (n – 1).m e i = j + m

SE (i - 1)mod(m) = 0

Ai,j = 0

SENÃO

Ai,j = - λ

PARA j=1 até n.m-1 e i = j + 1

Ai,j = -1

PARA j=1 ,.., nm e i=j

Ai,j = μ

PARA j=2 ,.., nm e i=j-1

Ai,j = -1

PARA j=1 +m até n.m e i = j - m

SE i mod(m) = 0

Ai,j = 0

SENÃO

Ai,j = - λ

Resolva sistema linear Aw = b

As equações simplificadas resultantes da aplicação de métodos numéricos podem chegar ao número de milhões, tornando a solução dos sistemas muito demorados em caso de processamento linear, portanto se torna necessário a utilização de sistemas computacionais mais avançados com a utilização de memória distribuída e compartilhada para a solução do problema através de processamento em paralelo

Particionamento para Diferenças Finitas: Calcula os limites das partições do domínio dos dados para cada processo.

ENTRADA: maxprocs: número de processos;

idproc: identificação do processo;

n: quantidade total de elementos.

SAÍDA:nidproc: número de elementos de dados da partição;

inicidproc: índice do elemento inicial da partição.

$$nidproc = \frac{n}{maxprocs}$$

inicidproc = idproc x nidproc

SE idproc< n mod(maxprocs)

Inicidproc = inicidproc + idproc

Nidproc = nidproc + 1

SENÃO

Inicidproc = inicidproc + n mod(maxprocs)

5.4) Quarta etapa do plano de trabalho:

O código legado, escrito em FORTRAN, e desenvolvido para o modelo computacional, é dividido em 14 módulos que contém uma variedade de funcionalidades matemáticas e físicas. O mesmo recebeu otimizações de paralelização de diretivas OpenMP. Com o uso das ferramentas do suite da INTEL Parallel Studio, como o VTune Amplifier, Thread Advisor e o Parallel Advisor, foi possível identificar os hotspots do código, possibilitando assim uma análise mais apurada do código dentro do módulo com o maior gasto computacional chamado transporte.F90 onde se encontra o problema numérico.

Foi possível atingir um ganho de 74x na execução do código no co-processador de arquitetura Many Integrated Core (MIC) Xeon-PHI KNL utilizando estratégias de otimização em OpenMP, porém, foi identificado via VTune que grande parte do tempo de execução da rotina era gasto em operações de sincronização entre threads. Foi aplicada uma estratégia para mudar o padrão de escalonamento das threads geradas pela API do compilador através do modelo de programação de memória compartilhada padrão OpenMP onde foi possível reduzir o spintime(sincronismo entre threads) de forma considerável reduzindo assim o CPI rate que é o índice de latência presente na execução do código, gerando também um ganho de desempenho de 27x, levando em consideração o tempo de execução do código com afinidade de threads Balanced, comparado ao código naive, executado de forma serial.

Apesar de realizar cálculos de dados 3D, o código legado foi criado inicialmente para realizar cálculos em 2d, levando assim a uma programação fora das boas práticas para a paralelização e causando um impacto negativo nos esforços de otimização. Em inúmeras áreas do código podemos ver condicionais aninhadas, tanto em outras condicionais, quanto em loops durante a execução do código, como também podemos ver inúmeras escritas em disco, também em loops, durante a execução. Os estudos que estão sendo desenvolvidos nesta etapa, são para encontrar o caminho de reestruturação total do código, voltado a programação seguindo as boas práticas da paralelização e focando a escalabilidade , desta forma, atingindo ganhos de desempenho ainda maiores comparados aos que puderam ser obtidos apenas utilizando estratégias, em OpenMP, no código legado.

6. Conclusão

Os estudos realizados são a base dos conhecimentos de matemática e de programação necessários para o desenvolvimento de análises e soluções de equações diferencias ordinárias e parciais através de métodos numéricos. A compreensão da solução de um modelo numérico discretizado é necessária para o entendimento das técnicas de solução através de processamento em paralelo.

Apesar de realizar cálculos de dados 3D, o código legado foi criado inicialmente para realizar cálculos em 2d, levando assim a uma programação fora das boas práticas para a paralelização e causando um impacto negativo nos esforços de otimização. Os estudos que estão sendo desenvolvidos nesta etapa, são para desenvolver a reestruturação total do código baseado em uma metodologia de programação orientada a objeto para ambientes de alto desempenho.

7. Referências Bibliografias

BURDEN, R. L.;FAIRES, J. D..**Análise Numérica**. Editora Thomson, (2003), p.605-614.

CUNHA, R. D. da. **Computação Paralela**: Uma Breve Introdução, XX CNMAC, (1997) p.5-20.

SANTOS, Juliano Deividy Braga. **Métodos de diferenças finitas de alta ordem para a equação da onda**, xv. Programa de Pós-Graduação de Modelagem Computacional (Dissertação), Laboratório Nacional de Computação Científica, Petrópolis, 2016, 134p.

OSTHOFF, Carla. **Introdução a Programação Paralela/*OpenMP***. Laboratório Nacional de Computação Científica, Petrópolis.

Relatório de Atividades: Inicialização de Fluidos para Animação Computacional

Bolsista: Allan Carlos Amaral Ribeiro Orientador: Gilson Antonio Giraldi
Tipo da Bolsa: PIBITI-LNCC
Período do Relatório: 01/08/2018 a 31/07/2019

18 de julho de 2019

1 Objetivos

Este projeto tem como objetivo estudar e desenvolver novos métodos para visualização científica de campos tensoriais, abordando a animação de fluidos. Desta forma, nesta etapa, o principal foco do trabalho foi desenvolver um modelo de projeção anisotrópico robusto que viabilizasse o uso de campos tensoriais como forma efetiva de controlar o fluido.

2 Introdução

Métodos para controlar simulações de fluido têm sido foco de pesquisas há mais de uma década, tendo computação gráfica como uma das inúmeras aplicações. Métodos para animação de fluidos são computacionalmente custosos dada a natureza inerentemente complexa da mecânica de fluidos, tornando ainda mais difícil a criação de métodos robustos para controle da simulação sem perder a plausibilidade física [11].

O trabalho de [11] foi um dos primeiros a abordar o problema de controle de fluidos, usando um conjunto de imagens de frames definidos pelo usuário para guiar a simulação. Em [5], o fluido é controlado através de uma função de distância, que é usada para gerar uma força externa. Há também métodos iterativos, como o de esculpir o fluido [4], no qual o usuário pode editar ativamente a produção da animação. Enquanto esses trabalhos focam o estado final do fluido, há os que buscam controlá-lo durante toda a simulação. A maioria usa forças externas para restringir caminhos pelo qual o fluido escoa. Em [3], uma força de controle é projetada de forma que o campo de velocidade converge para o fluxo especificado. Outros métodos, como [7] e [9], são baseados na deformação da malha da simulação. Interpolação de campos é outra alternativa, tendo sido empregada em [8] de forma a gerar animações de fluidos incompressíveis.

Nesse projeto, temos o interesse em desenvolver meios naturais de controlar fluidos, tanto em relação ao caminho, quanto ao estado final. Para isso, foram utilizados campos tensoriais como parte inerente do meio. Foi empregada uma abordagem baseada em malha, na qual define-se um campo tensorial que permite desviar o fluido, alterando os vetores de velocidades de forma a atender os objetivos da animação. Essa abordagem apresentada em [6], usa tensores simétricos semidefinidos positivos de segunda ordem. O método é uma adaptação de [10], em que o campo tensorial é inserido na formulação matemática de forma que ele altera localmente o momento linear do fluido. Cada uma das etapas básicas da simulação (advecção, difusão e projeção) foi alterada para levar em consideração a influência do tensor, refletindo as equações adaptadas.

Tendo o trabalho de [6] como ponto de partida, durante o período deste relatório, foi desenvolvido uma adaptação do método de projeção de velocidades para campos tensoriais definidos positivos aplicando a decomposição de Helmholtz anisotrópica. Além disso, também foi definida uma formulação contínua para as equações de Navier-Stokes utilizando um campo tensorial para transportar o fluido.

3 Metodologia Proposta

O principal fato explorado nessa etapa do trabalho é a existência e unicidade da decomposição anisotrópica de Helmholtz [2]. Dessa forma foi definido um método de projeção que pode ser usado em simulações de fluido incompressíveis com transporte anisotrópico. Todos os métodos relatados a seguir foram implementados e aprimorados pelo bolsista.

3.1 Decomposição Anisotrópica de Helmholtz

Seja $\mathbf{S}$ um tensor positivo definido de segunda ordem em $\mathbb{R}^3$, $s_1 \geq s_2 \geq s_3 > 0$ seus autovalores e $\{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3\}$ seus correspondentes autovetores. Como $\{s_1, s_2, s_3\}$ são autovalores correspondentes aos autovetores ortonormais $\{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3\}$, o $\mathbf{S}$-gradiente $\nabla_{\mathbf{S}}$ é definido como

$$\nabla_{\mathbf{S}} = \mathbf{S} \cdot \nabla = \begin{bmatrix} s_1 \frac{\partial}{\partial s_1} \\ s_2 \frac{\partial}{\partial s_2} \\ s_3 \frac{\partial}{\partial s_3} \end{bmatrix}. \tag{1}$$

Como resultado, o gradiente na base canônica é escalado e rotacionado por $\mathbf{S}$, assim as características direcionais do meio anisotrópico descrito por $\mathbf{S}$ são incorporados ao operador $\nabla_{\mathbf{S}}$. O $\mathbf{S}$-divergente e a $\mathbf{S}$-rotacional de um campo vetorial contínuo e diferenciável $\mathbf{f} : \Omega \subset \mathbb{R}^3 \to \mathbb{R}^3$ são definidos como:

$$\nabla_{\mathbf{S}} \cdot \mathbf{f}(\mathbf{r}) = (\mathbf{S} \cdot \nabla) \cdot \mathbf{f}(\mathbf{r}),$$

e

$$\nabla_{\mathbf{S}} \times \mathbf{f}(\mathbf{r}) = (\mathbf{S} \cdot \nabla) \times \mathbf{f}(\mathbf{r}),$$

respectivamente, onde $\mathbf{r} \in \Omega$. Agora, considere outro tensor positivo definido de segunda ordem $\mathbf{T}$ com o $\mathbf{T}$-gradiente

$$\nabla_{\mathbf{T}} = \begin{bmatrix} t_1 \frac{\partial}{\partial t_1} \\ t_2 \frac{\partial}{\partial t_2} \\ t_3 \frac{\partial}{\partial t_3} \end{bmatrix},$$

que leva em consideração as características de $\mathbf{T}$. Define-se o $\mathbf{ST}$-Laplaciano como

$$\nabla_{\mathbf{S}} \cdot \nabla_{\mathbf{T}} = (\mathbf{S} \cdot \nabla) \cdot (\mathbf{T} \cdot \nabla) = \nabla \cdot \mathbf{S} \cdot \mathbf{T} \cdot \nabla = \left(\sum_{i=1}^{3} s_i \mathbf{s}_i \frac{\partial}{\partial s_i} \right) \cdot \left(\sum_{j=1}^{3} t_j \mathbf{t}_j \frac{\partial}{\partial t_j} \right).$$

Se $\mathbf{S} = \mathbf{T}$, a expressão acima pode ser simplificada para

$$\nabla_{\mathbf{S}} \cdot \nabla_{\mathbf{S}} = s_1^2 \frac{\partial^2}{\partial s_1^2} + s_2^2 \frac{\partial^2}{\partial s_2^2} + s_3^2 \frac{\partial^2}{\partial s_3^2},$$

e se $s_1 = s_2 = s_3$, a equação pode ser simplificada para o Laplaciano isotrópico.

A versão anisotrópica do teorema de decomposição de Helmholtz assume a forma

$$\mathbf{f}(\mathbf{r}) = \nabla_{\mathbf{T}} \phi(\mathbf{r}) + \nabla_{\mathbf{S}} \times \mathbf{a}(\mathbf{r}), \tag{2}$$

onde o **T**-gradiente de ϕ é um campo **T**-irrotacional e a **S**-rotacional de **a** é um campo **S**-solenoidal. É possível encontrar o campo escalar $\phi(\mathbf{r}) = \nabla_{\mathbf{S}} \cdot \mathbf{h}(\mathbf{r})$ e o campo vetorial $\mathbf{a}(\mathbf{r}) = -\nabla_{\mathbf{T}} \times \mathbf{h}(\mathbf{r})$ resolvendo a equação **ST**-Poisson

$$\nabla_{\mathbf{S}} \cdot \nabla_{\mathbf{T}} \mathbf{h}(\mathbf{r}) = \mathbf{f}(\mathbf{r}). \tag{3}$$

Dassios e Lindell [2] provaram que a decomposição dada pela Equação 2, satisfazendo

$$\nabla_{\mathbf{T}} \cdot \mathbf{a} = -\nabla_{\mathbf{T}} \cdot \nabla_{\mathbf{T}} \ \times \ \mathbf{h}(\mathbf{r}) = 0,$$

é única. Como consequência, a hipótese sugerida em [6], que uma projeção anisotrópica pode ser usada para controlar a dinâmica de fluidos é plausível e esse foi o foco deste trabalho.

3.2 Deflexão tensorial do fluido

Nesse trabalho, a projeção anisotrópica em $\mathbb{R}^3$ requer tensores simétricos positivo definidos na forma:

$$\mathbf{T} = \sum_{i=1}^{3} t_i \mathbf{t}_i \mathbf{t}_i{}^T. \tag{4}$$

Chamados de tensores de orientação, estes podem ser utilizados para direcionar o fluido em caminhos específicos. O objetivo é usar o campo tensorial para controlar as simulações. A velocidade do fluido é transformada pelo tensor, defletindo-o localmente, podendo aumentar ou diminuir o momento.

A simulação implementada segue a abordagem clássica definida em [10], com densidade e velocidade discretizadas utilizando a representação Euleriana. O campo tensorial é representado em uma grade, onde cada célula armazena um tensor além dos valores referentes ao fluido, como velocidade e densidade. A principal ideia por trás da deflexão de fluidos com campos tensoriais, apresentada em [6], é que cada tensor deve funcionar como um modulador local das velocidades. Isso é alcançado através da seguinte equação diferencial:

$$\frac{\partial \mathbf{u}}{\partial t} = (\mathbf{T} - \beta \mathbf{I})\mathbf{u}, \tag{5}$$

onde β é um fator de escala introduzido por questões dimensionais.

A equação pode ser adicionada ao sistema de equação de Navier-Stokes de duas formas. Ao ser colocada na advecção, ela pode ser usada como uma restrição ao campo de velocidades, tratando o campo tensorial como parte inerente ao meio. Essa abordagem foi a adotada em [6] para mostrar como o campo tensorial pode controlar a advecção do fluido. Neste trabalho ela é usada como força externa, de forma a focar na etapa de projeção anisotrópica.

3.3 Método proposto

3.3.1 Projeção Anisotrópica de Helmholtz

Dado um campo de velocidades arbitrário $\mathbf{u}(\mathbf{r})$ obtido através da solução da equação de conservação de momento, a operação de projeção é o processo de encontrar um campo vetorial $\mathbf{w}(\mathbf{r}) = \mathrm{P}(\mathbf{u}(\mathbf{r}))$, que satisfaça a condição de incompressibilidade $\nabla \cdot \mathbf{w} = 0$. Muitos modelos de simulação forçam a incompressibilidade usando a decomposição clássica de Helmholtz-Hodge [1].

O campo anisotrópico, consequentemente, necessita de um método de projeção próprio. Como discutido em [6], a decomposição isotrópica de Helmholtz-Hodge pode violar as restrições locais, representadas pelo tensor. A projeção deve ter como resultado um campo vetorial que não só é livre de divergente, mas também atende as restrições de troca impostas pelo campo tensorial. Considere o **S**-divergente e o

$\mathbf{T}$-divergente aplicado a um campo vetorial arbitrário $\mathbf{f}(\mathbf{r})$ na Equação (2):

$$\nabla_{\mathbf{T}} \cdot \mathbf{f}(\mathbf{r}) = \nabla_{\mathbf{T}} \cdot \nabla_{\mathbf{T}} \phi(\mathbf{r}) + \nabla_{\mathbf{T}} \cdot \nabla_{\mathbf{S}} \times \mathbf{a}(\mathbf{r}), \tag{6}$$

$$\nabla_{\mathbf{S}} \cdot \mathbf{f}(\mathbf{r}) = \nabla_{\mathbf{S}} \cdot \nabla_{\mathbf{T}} \phi(\mathbf{r}) + \nabla_{\mathbf{S}} \cdot \nabla_{\mathbf{S}} \times \mathbf{a}(\mathbf{r}). \tag{7}$$

O objetivo é encontrar um campo vetorial livre de divergente através destas decomposições. Note que apenas a Equação 7 pode nos dar um campo vetorial livre de divergente. Se ambos $\mathbf{S} = \mathbf{I}$ e $\mathbf{T} = \mathbf{I}$, a decomposição é $\nabla \cdot \mathbf{f}(\mathbf{r}) = \nabla \cdot \nabla\phi(\mathbf{r}) + \nabla \cdot \nabla \times \mathbf{a}(\mathbf{r}) = \nabla \cdot \nabla\phi(\mathbf{r})$, e o campo vetorial do termo $\nabla \times \mathbf{a}(\mathbf{r})$ é livre de divergente em respeito a base canônica. Esta é a decomposição isotrópica de Helmholtz-Hodge.

Nosso interesse é usar campos tensoriais de forma que $\mathbf{S} \neq \mathbf{I}$ ou $\mathbf{T} \neq \mathbf{I}$. Nesse projeto, foi procurado soluções usando apenas um campo tensorial diferente da identidade. Primeiramente, considere $\mathbf{T} = \mathbf{I}$ e $\mathbf{S} \neq \mathbf{I}$, reduzindo a Equação 7 para:

$$\nabla_{\mathbf{S}} \cdot \mathbf{f}(\mathbf{r}) = \nabla_{\mathbf{S}} \cdot \nabla\phi(\mathbf{r}) + \nabla_{\mathbf{S}} \cdot \nabla_{\mathbf{S}} \times \mathbf{a}(\mathbf{r}) = \nabla_{\mathbf{S}} \cdot \nabla\phi(\mathbf{r}),$$

que pode gerar um campo vetorial livre de divergente $\nabla_{\mathbf{S}} \times \mathbf{a}(\mathbf{r})$. No entanto, esse campo é solenoidal em respeito ao campo tensorial $\mathbf{S}(\mathbf{r})$. De forma alternativa, ao definir $\mathbf{T} \neq \mathbf{I}$ e $\mathbf{S} = \mathbf{I}$, obtém-se

$$\nabla \cdot \mathbf{f}(\mathbf{r}) = \nabla \cdot \nabla_{\mathbf{T}} \phi(\mathbf{r}) + \nabla \cdot \nabla \times \mathbf{a}(\mathbf{r}) = \nabla \cdot \nabla_{\mathbf{T}} \phi(\mathbf{r}).$$

que pode gerar um campo livre de divergente $\nabla \times \mathbf{a}(\mathbf{r})$, solenoidal em respeito à base canônica. Desta forma, a projeção anisotrópica $\mathrm{P}_{\mathbf{T}}(\cdot)$ de um campo vetorial arbitrário $\mathbf{f}(\mathbf{r})$ restringido pelo campo tensorial $\mathbf{T}(\mathbf{r})$ é obtida solucionando a equação $\mathbf{T}$-Poisson

$$\nabla \cdot \nabla_{\mathbf{T}} \phi(\mathbf{r}) = \nabla \cdot \mathbf{f}(\mathbf{r}), \tag{8}$$

de modo a obter o campo vetorial projetado da Equação (2) como:

$$\mathrm{P}_{\mathbf{T}}(\mathbf{f}(\mathbf{r})) = \mathbf{f}(\mathbf{r}) - \nabla_{\mathbf{T}} \phi(\mathbf{r}), \tag{9}$$

que é livre de divergente em respeito a base canônica e sujeito à anisotropia de $\mathbf{T}$ e adequado para advecção incompressível em coordenadas cartesianas.

3.3.2 Equações Anisotrópicas para Fluidos Incompressíveis

Como em [6], mantém-se a ideia de que o tensor deflete localmente as velocidades, alterando o momento em um processo que pode ser parametrizado para atingir o controle desejado.

Seja $\mathbf{T}$ um tensor simétrico positivo definido, de um campo tensorial estacionário $\mathbf{T}(\mathbf{r})$, com autovalores $\{t_1, t_2, t_3\}$ e autovetores $\{\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3\}$.

Como usual, as quantidades dinâmica envolvidas são a velocidade do fluido $\mathbf{u}$, densidade ρ, pressão p, e um campo de forca externa $\mathbf{g}$. A principal ideia é modificar as equações de momento e massa usando a direção do campo tensorial e o termo de forca externa na expressão (5), de modo a aplicar a projeção anisotrópica ao invés da tradicional usada em [6]. Deste modo, a decomposição anisotrópica de Helmholtz torna-se parte central da formulação contínua, definida pelas seguintes expressões:

$$\frac{\partial \mathbf{u}}{\partial t} = -\mathbf{u} \cdot \nabla\mathbf{u} - \frac{1}{\rho}\nabla_{\mathbf{T}} p + \nu \nabla \cdot \nabla_{\mathbf{T}} \mathbf{u} + \mathbf{T}\mathbf{u} - \beta\mathbf{u} + \mathbf{g}, \tag{10}$$

$$\nabla \cdot \mathbf{u} = 0, \tag{11}$$

$$\frac{\partial \rho}{\partial t} = -\mathbf{u} \cdot \nabla\rho + \theta \nabla \cdot \nabla_{\mathbf{T}} \rho - \alpha\rho + s, \tag{12}$$

com condições iniciais $\mathbf{u}(\mathbf{r}, 0) = \mathbf{0}$ e $\rho(\mathbf{r}, 0)$ arbitrário, para $\mathbf{r} \in \Omega \subset \mathbb{R}^3$, onde o $\mathbf{T}$-gradiente $\nabla_{\mathbf{T}}$ é definido na expressão (1), a viscosidade ν e o parâmetro θ controlam a influencia de termos difusivos que são

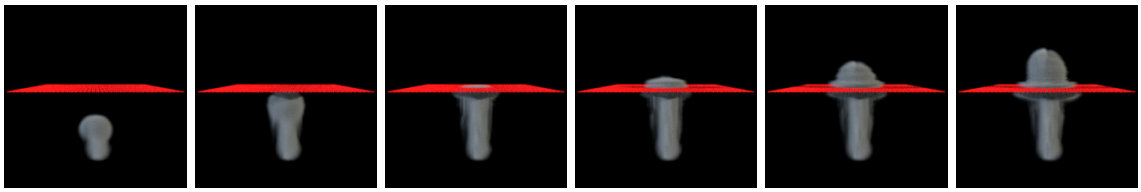


Figura 1: Simulação usando uma lâmina de tensores planares. Frames 30, 55, 80, 105, 130 e 155, respectivamente. O fluido atinge a camada de tensores planares (em vermelho) no frame 55, mas só consegue atravessa-lo, com momento já reduzido, depois do frame 130.

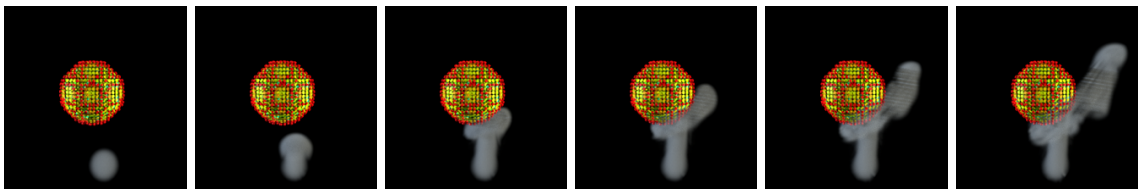


Figura 2: Simulação usando campo tensorial da esfera. Com o maior autovalor menor que um na casca da esfera, parte da densidade é retida na região da casca e parte é defletida.

afetados pelo tensor $\mathbf{T}$, para a velocidade e densidade, respectivamente. Além disso, s representa a fonte de densidade e o parâmetro real α escala a taxa de matéria que desaparece ao longo da simulação.

Esta formulação é usada para demonstrar como a projeção anisotrópica pode beneficiar o controle do fluido, mas o método desenvolvido pode ser adaptado para outros modelos de simulação.

4 Resultados e Discussão

Para avaliar a validade do novo método proposto, foram desenvolvidos diferentes casos de teste. Na construção dos campos tensoriais, utilizou-se a combinação linear de um conjunto limitado de tensores positivos definidos em $\mathbb{R}^3$ (com autovalores $\lambda_1 \geq \lambda_2 \geq \lambda_3$), classificados como planares($\lambda_1 > 0, \lambda_1 \approx \lambda_2 \gg \lambda_3$) e isotrópicos($\lambda_1 > 0, \lambda_1 \approx \lambda_2 \approx \lambda_3$)

Primeiramente foram utilizados campos tensoriais como lâminas de tensores planares. Os experimentos consistiam em uma fonte pontual de força externa de forma que o fluido fosse lançado em direção às lâminas de tensores, a fim de testar sua capacidade de contenção do fluido. A Figura 1 mostra um dos exemplos com a lâmina de tensores planares. Tais experimentos mostram que a projeção anisotrópica pôde isolar quase que completamente duas ou mais secções do volume, sendo necessário que o fluido tenha grande quantidade de momento para atravessá-la.

Em outros casos, uma esfera foi utilizada para dividir o domínio em três regiões, a região interna e externa à esfera foram preenchidas com tensores isotrópicos, enquanto os voxels sobrepostos pela superfície da esfera foram preenchidos com tensores planares orientados pela sua normal. Nestes testes aplicava-se diferentes tipos de forças externas sobre a casca e observava-se o comportamento do fluido. Os experimentos com a esfera mostraram que, utilizando a projeção anisotrópica, é possível utilizar o campo tensorial tanto para conter, quanto para espalhar o fluido em regiões de interesse.

Em vista dos resultados, o método de projeção anisotrópica se mostrou de grande utilidade no controle do fluido, permitindo maior controle sobre a simulação utilizando campo tensoriais. A análise detalhada do modelo matemático desenvolvido neste trabalho, junto com o desenvolvimento da solução numérica estão sendo relatados em um artigo a ser submetido à revista Applied Mathematics and Computation, que é A1 em Ciência da Computação.

5 Conclusões

O presente trabalho se mostra de grande importância para o uso de campos tensoriais como meio de controle de simulações de fluido, uma vez que o método de projeção anisotrópica ajuda a garantir as restrições ao fluxo impostas localmente pelos tensores. Este trabalho também mostra como tal método pode ser integrado à um modelo de simulação de fluidos alterando a formulação contínua.

Os resultados mostraram que mesmo campos tensoriais simples podem ser utilizados para animações mais elaboradas. Este modelo também torna viável o uso de campos tensoriais mais complexos, provendo maior flexibilidade e controle à métodos de inicialização de fluidos para animação computacional utilizando campos tensoriais.

Como produto desse projeto, um artigo, já em fase de consolidação de resultados, está sendo confeccionado e será submetido à revista Applied Mathematics and Computation.

Referências

[1] Robert Bridson. *Fluid simulation for computer graphics*. AK Peters/CRC Press, 2015.

[2] George Dassios and Ismo V Lindell. Uniqueness and reconstruction for the anisotropic helmholtz decomposition. *Journal of Physics A: Mathematical and General*, 35(24):5139, 2002.

[3] Yootai Kim, Raghu Machiraju, and David Thompson. Path-based control of smoke simulations. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 33–42. Eurographics Association, 2006.

[4] Pierre-Luc Manteaux, Ulysse Vimont, Chris Wojtan, Damien Rohmer, and Marie-Paule Cani. Space-time sculpting of liquid animation. In *Proceedings of the 9th International Conference on Motion in Games*, pages 61–71. ACM, 2016.

[5] Marcelo Caniato Renhe, Antonio Oliveira, Cláudio Esperança, and Ricardo Marroquim. Enhanced target driven smoke morphing. In *Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on*, pages 213–220. IEEE, 2012.

[6] Marcelo Caniato Renhe, Marcelo Bernardes Vieira, and Claudio Esperança. A stable tensor-based method for controlled fluid simulations. *Applied Mathematics and Computation*, 343:195–213, 2019.

[7] Syuhei Sato, Yoshinori Dobashi, Kei Iwasaki, Tsuyoshi Yamamoto, and Tomoyuki Nishita. Deformation of 2d flow fields using stream functions. In *SIGGRAPH Asia 2014 Technical Briefs*, page 4. ACM, 2014.

[8] Syuhei Sato, Yoshinori Dobashi, and Tomoyuki Nishita. A combining method of fluid animations by interpolating flow fields. In *SIGGRAPH ASIA 2016 Technical Briefs*, page 4. ACM, 2016.

[9] Syuhei Sato, Yoshinori Dobashi, Yonghao Yue, Kei Iwasaki, and Tomoyuki Nishita. Incompressibility-preserving deformation for fluid flows using vector potentials. *The Visual Computer*, 31(6-8):959–965, 2015.

[10] Jos Stam. Stable fluids. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 121–128. ACM Press/Addison-Wesley Publishing Co., 1999.

[11] Adrien Treuille, Antoine McNamara, Zoran Popović, and Jos Stam. Keyframe control of smoke simulations. *ACM Transactions on Graphics (TOG)*, 22(3):716–723, 2003.

1. Dados Gerais

Aplicação das ferramentas Intel Parallel Studio para modernização de código para métodos numéricos de diferenças finitas para solução de equações diferenciais parciais em arquitetura Intel Haswel/Broadwell.

Bolsista: Gabriel Pinheiro da Costa

Orientadora: Carla Osthoff

Co-orientador: Frederico Luís Cabral

Bolsa de Iniciação Tecnológica pelo CNPq

Esse relatório cobre as atividades realizadas entre agosto de 2018 até julho de 2019

2. Objetivos

O objetivo é o desenvolvimento do conhecimento na área de processamento de alto desempenho para avaliar e otimizar o desempenho de uma aplicação de computação científica em uma máquina de arquitetura paralela haswell/broadwell, com o uso das ferramentas de análise de código

3.Introdução

O trabalho almeja realizar um estudo das oportunidades de otimização de código para a solução de equações diferenciais parciais através das ferramentas do Intel Parallel Studio.

Os métodos numéricos destinados à solução dessas equações são contemplados pela possibilidade vasta do uso de ferramentas e técnicas cada vez mais efetivas, que buscam aumentar o desempenho dessas aplicações, reduzindo o tempo de execução e permitindo entradas de dados de maior magnitude.

No contexto de otimização, ganha forte destaque a técnica de paralelização, que permite que os dados sejam processados simultaneamente por mais de uma unidade de processamento, sem comprometer a qualidade do resultado final, e reduzindo o tempo necessário para chegar até esse resultado.

As ferramentas do Intel Parallel Studio trabalham com informações do comportamento das instruções do código durante a compilação e a execução. A partir dessa gama de informações é possível extrair conclusões acerca da aplicação e seu

desempenho atual, e portanto nortear um plano de otimização para essa aplicação, buscando um aproveitamento ótimo do hardware utilizado.

4. Material e Métodos ou Metodologia

Para a realização dos testes das diferentes implementações do método HOPMOC foram usadas uma máquina contendo dois sockets Intel Xeon Platinum 8160 (SKL) @ 2.1GHz, com 24 cores por socket e 2 threads por core, e uma máquina contendo um acelerador Intel Xeon Phi KnightsLanding (KNL) com 68 cores de 1.4 GHz e 4 threads por core.

5. Resultados e Discussão

A forma mais simples de se realizar a paralelização do método numérico estudado na pesquisa, o HOPMOC, é através da inserção de diretivas padrão de paralelização da interface OpenMP. Essas diretivas de compilador permitem uma paralelização automática do código por parte da interface, distribuindo a carga das principais rotinas do código entre as threads disponíveis para execução.

```
1  #pragma simd
2  #pragma omp for
3  for (int i = head+1 ; i <= N-2 ; i+=2) {
4    ANNOTATE_ITERATION_TASK(loop_HOP_EXP_1);
5    U_old[i] = alfa*(U_new[i-1] + U_new[i+1]) +
         (1 - 2*alfa)*U_new[i];
6  }
7
8  #pragma omp single
9    head = (head+1)%2;
10
11 #pragma simd
12
13 #pragma omp for
14 for (int i = head+1; i <= N-2 ; i+=2) {
15   ANNOTATE_ITERATION_TASK(loop_HOP_IMP_1);
16   U_old[i] = (U_new[i] + alfa*U_old[i-1] + alfa
         *U_old[i+1])/(1+2*alfa);
17 }
18
19 #pragma omp single
20   head = (head+1)%2;
21
22 #pragma simd
23 #pragma omp for
24 for (int i = head+1 ; i <= N-2 ; i+=2) {
25   ANNOTATE_ITERATION_TASK(loop_HOP_EXP_2);
26   U_new[i] = alfa*(U_old[i-1] + U_old[i+1]) +
         (1 - 2*alfa)*U_old[i];
27 }
28
29 #pragma omp single
30   head = (head+1)%2;
31
32 #pragma simd
33 #pragma omp for
34 for (int i = head+1; i <= N-2 ; i+=2) {
35   ANNOTATE_ITERATION_TASK(loop_HOP_IMP_2);
36   U_new[i] = (U_old[i] + alfa*U_new[i-1] + alfa
         *U_new[i+1])/(1+2*alfa);
37 }
```

Todavia, o ganho de desempenho para essa primeira versão mais ingênua, nomeada Naive, é extremamente limitado por fatores como o tempo gasto em barreiras de sincronização entre as threads, tempo de alocação de threads a cada diretiva, e

desbalanceamento de carga. Á partir dessas limitações, foram desenvolvidas três estratégias distintas a fim de contornar os problemas identificados e permitir um ganho de desempenho menos limitado. As seções subsequentes expõem de forma breve as alternativas encontradas.

5.1 HOPMOC EWS-AdSynch baseado em OpenMP

Essa estratégia consiste em designar a distribuição de threads manualmente, ao contrário do alocação automática realizada pelo OpenMP através da diretiva *parallel for*, na versão Naive. Assim, no início do código, a malha é subdividia em um dado número de regiões correspondente ao número de threads disponíveis no momento, de sorte a designar permanentemente até o fim da execução, uma região específica a uma thread específica.

Nessa versão do HOPMOC, a sincronização de uma thread precisa ser feita, somente, com as threads adjacentes e não com todas as outras threads como na versão Naive. Para isso o HOPMOC EWS conta com um vetor booleano responsável por armazenar o status de cada thread (pronto para continuar ou não), o que permite uma sincronização explícita entre as threads vizinhas.

A grande vantagem dessa abordagem é a eliminação do tempo gasto em barreiras de sincronização, que antes abrangiam todas as threads envolvidas, e a redução do tempo de alocação de threads, uma vez que essa designação agora é feita somente uma vez no início do código. Também é possível afirmar que, através da alocação manual presente nessa versão, há um melhor balanceamento de carga, uma vez que a divisão matemática e explícita da malha de acordo com o número de threads disponíveis garante uma disposição homogênea da carga por linha de processamento paralelo.

5.2 HOPMOC baseado em MPI

Outra estratégia desenvolvida afim de contornar os problemas limitadores encontrados na versão Naive foi o uso de múltiplos processos para se alcançar o paralelismo, ao invés de threads, valendo-se do padrão MPI para promover a comunicação entre esses processos.

Nessa versão, a malha original é dividida em um certo número de malhas menores, de acordo com o número de processos disponíveis, de sorte a designar uma seção específica a um processo específico, em um processo semelhante à alocação explícita de threads do HOPMOC EWS. Cada processo deverá, portanto, alocar seu fragmento de malha correspondente à malha original.

A comunicação é feita apenas entre processos vizinhos, como no EWS, e é constituída da troca de dados referentes às fronteiras de cada submalha alocada por cada processo. Assim, um dado processo envia ao seu vizinho da esquerda a extremidade esquerda de sua malha e espera receber a extremidade direita dessa vizinho da esquerda, ao passo que envia ao seu vizinho da direita a extremidade direita de sua malha, esperando receber a extremidade esquerda desse vizinho da direita.

Essa abordagem apesar de trabalhar com processos e não com threads, possui vantagens semelhantes às do EWS: diminuição do tempo gasto em barreiras de sincronização, eliminação do tempo de alocação de threads e melhor balanceamento de carga.

5.3 HOPMOC híbrido baseado em MPI e OpenMP EWS

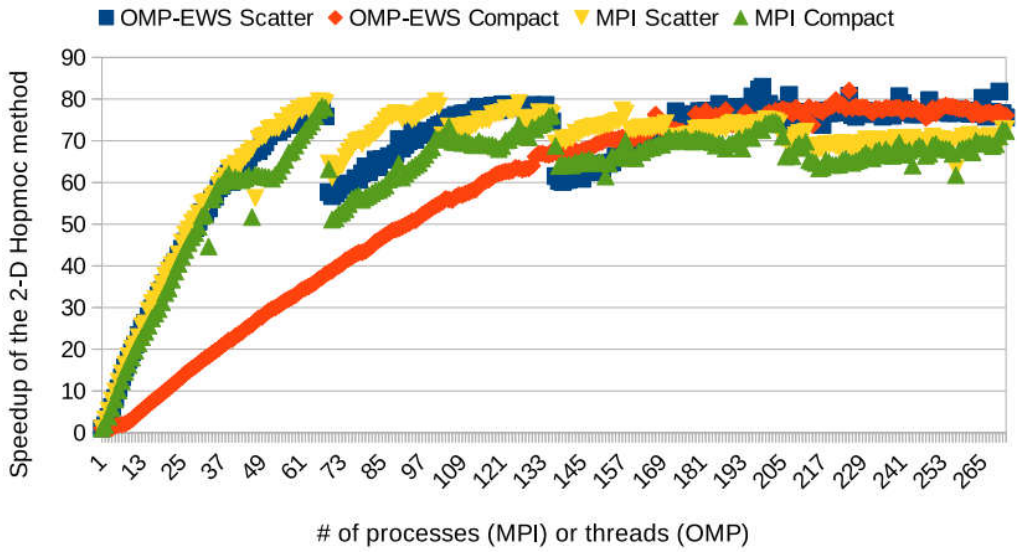
A última abordagem apresentada nesse resumo é um híbrido entre as duas abordagens mencionadas anteriormente: uma baseada em processos e comunicação MPI e a outra baseada na divisão explícita e sincronização manual de threads.

Assim, a malha original é dividida em submalhas menores que serão designadas a processos exatamente como descrito na seção anterior, mas além disso, no interior dos processos essas submalhas serão novamente divididas de forma explícita, e designadas a threads específicas sincronizadas por um vetor booleano, permanentemente até o fim da execução, como ocorre naturalmente na abordagem EWS.

Essa estratégia busca combinar o melhor das duas abordagens iniciais, tentando garantir dois níveis de paralelismo e a sincronização tanto pela troca de mensagens entre processos vizinhos através do MPI quanto pelo vetor booleano de estados das threads.

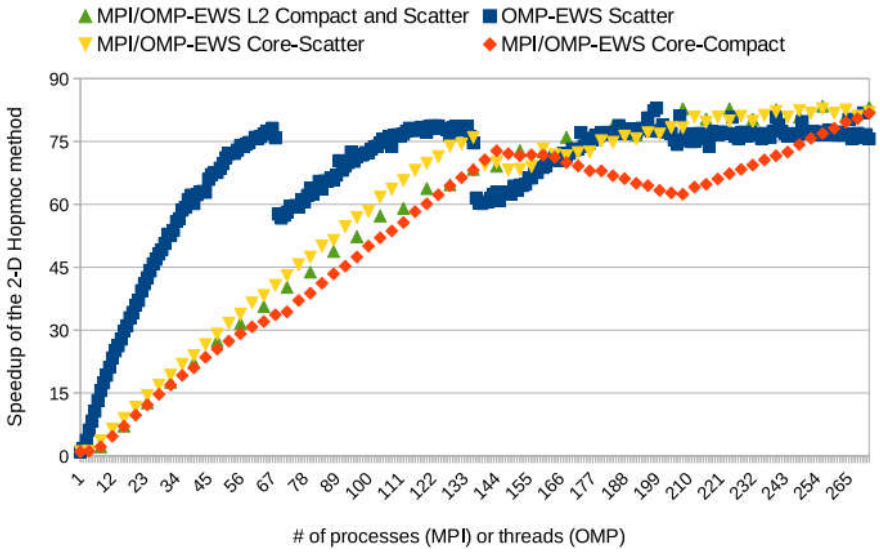
5.4 Comparação de desempenho

O gráfico abaixo aspresenta o speedup de quatro execuções distintas em uma máquina KNL: a versão EWS pura com as políticas de distribuição de threads Compact e Scatter, e a versão MPI pura com as políticas de distribuição de threads Compact e Scatter.



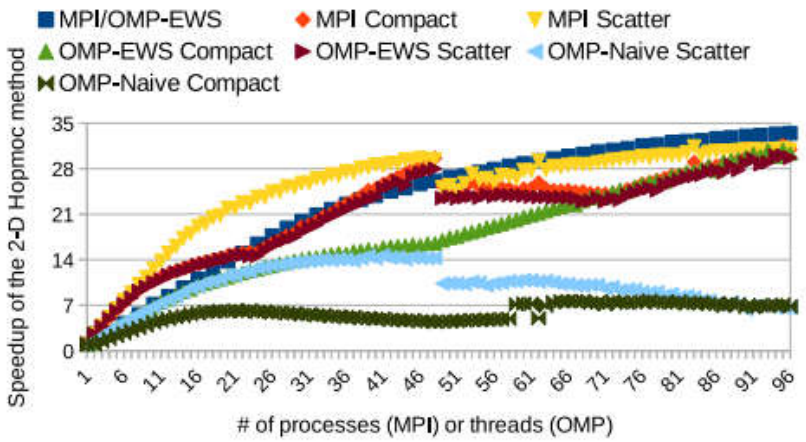
Pela figura é possível observar que a versão OMP-EWS com a política de distribuição de threads scatter conseguiu um speedup levemente superior com o número máximo de threads/processos.

Outro teste realizado na mesma máquina compara o speedup por número de threads de mais quatro execuções: o OMP-EWS scatter que alcançou o melhor desempenho no teste anterior, a versão híbrida (MPI/OMP-EWS) L2 com a política scatter/compact, híbrida Core com a política scatter e híbrida Core com a política compact.



A imagem permite concluir que a versão híbrida MPI/OMP-EWS L2 com a política compact ou scatter alcança um desempenho levemente superior às outras três execuções quando usada a quantidade máxima de threads/processos. O desempenho das três versões híbridas quando usadas todas as threads/processos é extremamente similar, uma vez que todos os cores e suas threads estão completamente ocupados, tornando a política de distribuição de threads, nesse instante, indiferente.

Execuções em uma máquina SKL permitem comparar o speedup da versão Naive com diferentes políticas de distribuição de threads em relação às versões OMP-EWS, MPI e MPI/OMP-EWS



6.Conclusões

Os testes realizados nas duas máquinas apontam para ganhos consideráveis de desempenho das estratégias desenvolvidas se comparadas à versão padrão Naive, podendo chegar a um speedup cerca de 5 vezes maior, como é o caso da versão híbrida MPI/OMP-EWS na SKL.

Esses ganhos demonstram a validade das estratégias desenvolvidas à partir das análises das ferramentas da Intel com o intuito de superar as contradições e limitações encontradas na versão mais ingênua do HOPMOC. Tanto as versões OMP-EWS e MPI e principalmente a versão híbrida MPI/OMP-EWS conseguem dar solução aos problemas de tempo gasto em barreira de sincronização de threads, tempo gasto com alocação de threads no núcleo do código e balanceamento de carga, satisfazendo as expectativas que impulsionaram sua criação.

7.Referências

RAUBER, Thomas; RÜNGER, Gudula. **Parallel programming: For multicore and cluster systems**. Springer Science & Business Media, 2010.

Cabral, F.L.; Osthoff, C.; Costa, G.P.; Gonzaga de Oliveira, S.L.; Brandão, D.; Kischinhevsky, M. *An OpenMP Implementation of the TVD Hopmoc Method Based on a Synchronization Mechanism Using Locks Between Adjacent Threads on Xeon Phi (TM) Accelerators*. Lecture Notes in Computer Science. Springer International Publishing, 2018, v. 3, p. 701-707. doi: 10.1007/978-3-319-93713-7_67

Cabral F.L.; Osthoff C.; Costa G.; Brandão D.; Oliveira S.L.G.; Kischinhevsky M. *Tuning up TVD HOPMOC method on Intel MIC Xeon Phi Architectures with Intel Parallel Studio Tools* Workshop on Applications for Multicore Architecture (WAMCA 2017) - International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD 2017) - Campinas - São Paulo - Brasil - Outubro de 2017.

TEIXEIRA, Thiago Daniel Quimas Simões. Estratégias de otimização para um Método Numérico para resolução de problemas de Convecção-Difusão. 2016. 72 f. Trabalho de Conclusão de Curso(Graduação em Sistema de Informação) - Universidade Estácio de Sá, Petrópolis, 2016.

LABORATÓRIO NACIONAL DE COMPUTAÇÃO CIENTÍFICA
LABORATÓRIO DE MODELAGEM EM HEMODINÂMICA
PROGRAMA INSTITUCIONAL DE BOLSAS DE INICIAÇÃO CIENTÍFICA

TÉCNICAS DE APRENDIZAGEM DE MÁQUINA NA SEGMENTAÇÃO DE IMAGENS MÉDICAS DE ULTRASSOM INTRAVASCULAR

Bolsista: Jefferson da silva Fernandes de Azevedo

Orientador: Pablo Javier Blanco

Petrópolis

Julho 2018 – Julho 2019

Sumário

1 OBJETIVOS 2
2 INTRODUÇÃO 2
3 METODOLOGIA 2
3.1 PREPARAÇÃO DOS DADOS PARA O TREINAMENTO 2
3.2 ELABORAÇÃO DO ALGORITMO DA REDE NEURAL 3
4 RESULTADOS E DISCUSSÃO 4
5 CONCLUSÃO 6
6 BIBLIOGRAFIA 6

1 OBJETIVOS

O trabalho consiste em estudar o uso de redes neurais artificiais e sua aplicação em problemas nos quais se conhece a física subjacente. O principal objetivo é criar um algoritmo capaz de destacar o lúmen e parede arterial em imagens médicas de IVUS (*Intravascular Ultrasound)* sem precisar de um médico especialista.

2 INTRODUÇÃO

Redes neurais artificiais são utilizadas quando não se conhece a função matemática que rege algum comportamento físico, ou quando esta função é matematicamente difícil de solucionar. Nestes casos, uma solução pode ser criar uma rede neural treinada em cima dos resultados conhecidos deste fenômeno. Uma rede neural é uma função matemática com um caráter genérico que pode ser adaptada a um fenômeno real. Esta adaptação depende dos valores de seus parâmetros (pesos e bias), que são determinados na fase de treinamento. Portanto, trazem possibilidades de elaborar ferramentas que beneficiem e ampliem novas abordagens para soluções de problemas.

Para o contexto abordado, a ultrassonografia intravascular, ou IVUS (sigla em inglês), é um exame que consiste em visualizar o interior de artérias do corpo humano. O exame é realizado ao introduzir um cateter pela artéria femoral e seu propósito é detectar as regiões do lúmen, parede arterial e o exterior da artéria. Ressalta-se que de acordo com a espessura da parede arterial, pode-se determinar regiões frágeis ao longo da artéria e se há acúmulo de gordura (aterosclerose).

3 METODOLOGIA

3.1 PREPARAÇÃO DOS DADOS PARA O TREINAMENTO

Em primeiro lugar há a necessidade de elaborar o material para treinamento da rede neural. O algoritmo a ser elaborado deverá receber imagens originais do IVUS (Figura 1 - a) e a saída será uma imagem indicando apenas 3 valores: 0 – exterior da artéria, 1 – parede arterial e 2 – lúmen (Figura 1 - b). Observa-se que a Figura 1- b é o objetivo do trabalho, porém deve ser gerada de forma automática.

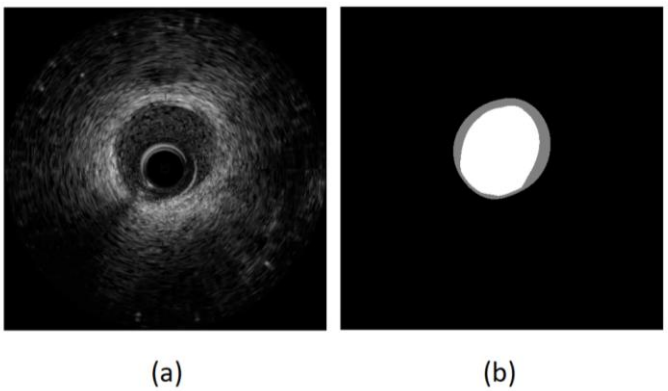


Figura 1: (a) Imagem Original de Ultrassonografia Intravascular.
(b) Representação das regiões de interesse na imagem: Branco (Lúmen), Cinza (Parede arterial) e Preto (Exterior da Artéria).

As imagens IVUS são figuras na escala de cinza com dimensão de 512x512 pixels. Portanto serão analisadas como matrizes bidimensionais de 512 linhas e 512 colunas com valores de 0 (preto) a 255 (branco).

Entretanto, ressalta-se que a Figura 1 – b não é um dado do problema. Para gerar esta imagem segmentada foi elaborado, como parte do presente trabalho, um algoritmo do *software* MATLAB no qual o especialista indica as regiões de delimitação das áreas de interesse, e o programa marca o interior desta região.

Assim, existe um esforço a ser realizado a fim de obter os dados considerados como dados de treinamento no contexto das redes neurais. As estruturas obtidas correspondem ao lúmen do vaso arterial, a parede da artéria e a região externa do vaso. Este algoritmo facilita a criação e depuração dos dados a serem utilizados no treinamento e, posteriormente, na validação da rede neural.

3.2 ELABORAÇÃO DO ALGORITMO DA REDE NEURAL

Para determinar o algoritmo será necessário ter conhecimento do funcionamento de redes neurais e de suas propriedades, como: função de ativação, função de perda, pesos e bias, épocas de treinamento e predição. Conhecer a influência de cada um destes itens é de grande importância para a função convergir para o valor desejado.

Para isso, desenvolveram-se redes baseadas em um artigo sobre redes neurais que utilizam equações diferenciais para controlarem a função de perda e melhorar o seu desempenho [1]. A função de perda, geralmente, é apenas a diferença entre a predição da rede e o resultado esperado. Ao acrescentar um termo à perda que está relacionado à equação diferencial que rege o sistema é garantido que além do resultado minimizar a diferença entre a predição e o previsto de acordo com o modelo matemático, os valores dos parâmetros terão uma penalidade caso desrespeitem o funcionamento físico do sistema.

Esta tarefa possui grande utilidade pois permite analisar, e manipular, os elementos de uma rede neural de forma individual. E, ao utilizar uma equação diferencial em que se conhece a solução, obtêm-se o controle da influência dos dados de entrada e se a predição está sendo feita de maneira correta.

A equação diferencial considerada neste estudo, assim como sua solução estão demonstradas abaixo, respectivamente. A solução analítica será utilizada para criar os dados de entrada e a equação diferencial será um parâmetro adicionado à função de perda.

Equação diferencial:

$$\frac{dy}{dt} + ay = 0$$

Solução analítica, dado $y(\alpha,0) = y0$:

$$y(\alpha, t) = y0.e^{-at}$$

A criação do algoritmo foi realizada baseada no código desenvolvido pelos autores citados anteriormente [1]. O intuito é estudar os comandos e raciocínios utilizados e, ao entender conceitualmente seu funcionamento, adaptar-se-á ao problema proposto como teste. Como plataforma de programação, utilizou-se a linguagem Python e a biblioteca TensorFlow.

Portanto, esta rede neural receberá um valor de alfa e t e se espera um resultado semelhante ao da solução analítica mencionada acima. Como parâmetro de melhoria na predição, será adicionado a equação diferencial na função de perda da rede.

4 RESULTADOS E DISCUSSÃO

O estudo do algoritmo e a convergência para a função determinada funcionou adequadamente. O sistema possui erro menor que 10% até um valor de α igual a 50% acima do máximo utilizado para treinar, e funciona bem para valores de t até 10 vezes maiores dos que foram informados na fase de treinamento. Quanto ao número de neurônios por camada, a primeira camada possui 2 neurônios (entradas da rede são: alfa e tempo), as camadas intermediárias são de 10 neurônios e a camada de saída possui um único neurônio (o valor da variável y).

O parâmetro mais crítico para a convergência foi a escolha dos valores iniciais dos pesos. A convergência ocorre somente ao selecionar valores iniciais de acordo com uma distribuição normal com desvio padrão dado pela fórmula abaixo.

$$desvio = \sqrt{\frac{2}{N°\ de\ neurônios\ na\ camada\ i + \ N°\ de\ neurônios\ na\ camada\ i + 1}}$$

Certamente vale salientar que apesar deste ter sido o fator mais crítico, não torna os outros obsoletos, um número baixo de camadas e de neurônios limita a não-linearidade da função, poucos dados de treinamento impossibilitam uma boa generalização e o acréscimo da equação diferencial na função perda fornece uma melhoria na convergência.

O método de avaliação da precisão da rede é dado pela equação abaixo:

$$acc = \left(1 - \left|1 - \frac{\sum y\_pred}{\sum y\_target}\right|\right).100\%$$

Considerando o intervalo de treinamento de $0 \leq \alpha \leq 10$ e $0 \leq t \leq 20$ os resultados obtidos podem ser vistos na tabela 1 e 2. Observe que a coluna Precisão é referente aos valores usados para treinamento da rede, e na coluna Validação são usados valores diferentes do treinamento, porém no mesmo intervalo de α e t. A coluna PINN (*Physics Informed Neural Network*) é referente se foi utilizado ou não a equação diferencial na perda da função.

N° camadas	Tempo de treinamento	Épocas	PINN	Perda Final	Precisão	Validação
5	28.25s	10,000	Sim	2.1054e-3	99.60%	99.39%
5	17.73s	10,000	Não	3.5775e-5	100,00%	99.67%
5	14.97s	5,000	Sim	6.7563e-3	99.97%	99.75%
10	48.56s	10,000	Sim	2.2116e-5	99,47%	98.90%
10	23.93s	10,000	Não	7.6924e-7	100,00%	99.54%
10	25.51s	5,000	Sim	6.4897e-3	99.94%	99.95%
15	1 min e 8.54s	10,000	Sim	4.4220e-5	99.93%	99.99%
15	30.65s	10,000	Não	3.1266e-7	100,00%	99.90%
15	33.49s	5,000	Sim	8.8432e-4	99.96%	99.99%
20	1 min e 33.43s	10,000	Sim	2.8268e-5	99.96%	99.87%
20	37.35s	10,000	Não	1.9500e-6	100,00%	99.98%
20	44.03s	5,000	Sim	1.5666e-2	99.34%	99.54%

Tabela 1: Grupo de treinamento com 1,000 dados.

N° camadas	Tempo de treinamento	Épocas	PINN	Perda Final	Precisão	Validação
5	23.48s	10,000	Sim	5.9895e-4	99.78%	99.59%
5	13.82	10,000	Não	1.6788e-5	100,00%	99.43%
5	12.28s	5,000	Sim	4.6196e-3	99.74%	98.96%
10	42.05s	10,000	Sim	2.5332e-4	89.39%	61.88%
10	21.91s	10,000	Não	1.0213e-7	99.99%	98.02%
10	21.67s	5,000	Sim	7.2733e-5	99.73%	99.74%
15	1 min e 3.44s	10,000	Sim	1.2367e-6	99.94%	96.86%
15	28.56s	10,000	Não	8.4361e-8	100,00%	98.61%
15	33.39s	5,000	Sim	2.6817e-4	99.65%	96.28%
20	1 min e 20.14s	10,000	Sim	4.3015e-3	99.56%	98.33%
20	36.53s	10,000	Não	9.1449e-7	99.99%	95.72%
20	41.62s	5,000	Sim	1.4218e-2	95.14%	77.67%

Tabela 2: Grupo de treinamento com 250 dados.

Na fase de validação, a rede neural começa a apresentar imprecisão no início da curva quando $\alpha > 15$ e diverge quando $\alpha < 0.7$ e $t > 200$. Os gráficos podem ser observados na Figura 3 a e b.

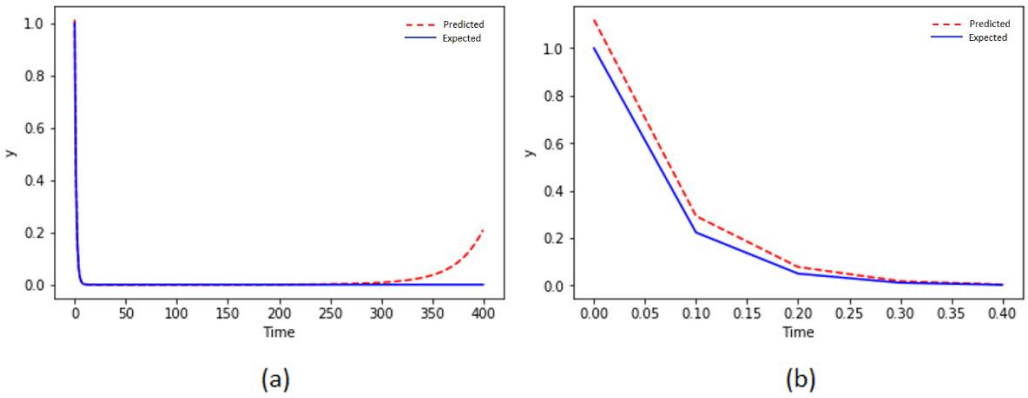


Figura 3: (a) $\alpha = 0.7$, divergência em $t > 200$ (b) $\alpha = 15$, imprecisão em $t < 0.3$

5 CONCLUSÃO

Para a segmentação das imagens médicas do IVUS, foi-se desenvolvido um *script* no MATLAB que fará a elaboração dos dados de treinamento em grande quantidade. E com isso possibilitou a criação do banco de dados que será usado no treinamento. Quanto ao desenvolvimento do algoritmo, a utilização de redes neurais baseadas em equações diferenciais foi uma ferramenta enriquecedora em prol da busca pela perícia da biblioteca TensorFlow aplicada em redes neurais. O programa elaborado possibilitou maior discernimento acerca do desenvolvimento de redes neurais e ajudou a visualizar a influência de cada um dos itens que compõem uma rede neural. Tais fatores são relevantes para o desenvolvimento do algoritmo definitivo.

6 BIBLIOGRAFIA

[1] - Raissi M., Perdikaris P. and Karniadakis G. E., *Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations*. University of Pennsylvania, Philadelphia, PA – 2017. Disponível em: https://github.com/maziarraissi/PINNs

RELATÓRIO DE ATIVIDADES

Título do Projeto: Gerência de Aplicações Científicas no Portal da Rede Nacional de Bioinformática (Bioinfo-Portal)

Nome do bolsista: Mayconn Luiz Bispo dos Santos

Nome do orientador:

D.Sc. Kary Ann del Carmen Ocaña Gautherot (Tecnologista Pleno – LABINFO/LNCC)

Nome dos coorientadores:

D.Sc. Antonio Tadeu Azevedo Gomes (Tecnologista Sênior – SINAPAD/LNCC)

B.Sc. Marcelo Monteiro Galheigo (SINAPAD/LNCC)

Tipo de bolsa: PIBIT

Período do relatório: 01/10/2018 a 31/07/2019

OBJETIVOS

O principal objetivo do Projeto de IC é implementar uma nova versão do Portal-Bioinfo (Portal de Bioinformática do LNCC) v2.0 mais segura e eficiente, adaptada e acoplado para ser executada no supercomputador Santos Dumont. Outros objetivos específicos do projeto são mapear os scripts de configuração e execução (xml e sh) do Portal, gerenciar os dados de proveniência (metadados de execução e domínio específico) das aplicações científicas de bioinformática do Portal e realizar um update das versões de PHP do Portal.

INTRODUÇÃO

O Projeto de Iniciação Científica (IC) se desenvolve no nível de estágio supervisionado sob coordenação da orientadora Kary Ocaña (LNCC) e com bolsa IC financiada pelo CNPq. O mesmo se enquadra nas pesquisas relacionadas aos projetos institucionais do LNCC, especificamente da Rede Nacional de Bioinformática (RNBio) e do Portal-Bioinfo (https://bioinfo.lncc.br/). O Portal-Bioinfo visa a execução em larga escala de aplicações de bioinformática usando recursos computacionais paralelos e distribuídos a fim de diminuir o grande tempo de processamento das execuções.

O Projeto de IC visa o desenvolvimento de uma versão do Portal-Bioinfo mais segura, eficiente e interativa. Dessa maneira, o projeto apresenta as seguintes atividades: (1) mapear scripts do Bioinfo-Portal e atualizar o código (PHP) para versões mais atuais; (2) acoplar tecnologias utilizados pela interface Web do portal (HTML, CSS, JavaScript, Bootstrap, Google charts, etc.) e (3) gerenciar e explorar os dados por meio do uso de Sistemas de Gerência de Banco de Dados (SGBD) e técnicas de aprendizado de máquina.

A atualização da nova versão do Portal-Bioinfo irá prover uma melhor acessibilidade do Portal e maior número de acessos no uso das aplicações por parte de usuários e a comunidade científica de Bioinformática no Brasil.

METODOLOGIA

Foram exploradas as tecnologias e acopladas na arquitetura do Portal, sob o presente projeto: PHP7, Google Charts, Reporting API V4, Html, Css, Mysql, Postgree, Bootstrap, Javascript, VPN-Linux, Angula JS, Eclipse, Sublime Text, Easy PHP, etc.

Foi necessário o estudo de comandos do SO (Sistema Operacional) utilizado para o ambiente de desenvolvimento das aplicações do projeto, no nosso caso, o sistema operacional LINUX.

RESULTADOS E DISCUSSÃO

Sobre as implementações no decorrer do estágio apresentamos três principais resultados: (1) a implementação de 10 novas aplicações no portal (status finalizado); (2) a implementação do Google Analytics (status finalizado) e (3) a atualização das versões PHP5 para PHP7 (status em andamento). 10 novas aplicações científicas (bcftools, BEAST2, bwa, bwa-aln, bwa-build, ExaML, ExaML-raxml, ExaML-parser, NxTrim, PartitionFinder) foram incluídas no Portal-Bioinfo, somando aproximadamente um total de 40 scripts (xml, PHP e sh). Esses scripts foram acoplados na versão de teste do Portal-Bioinfo, estão em funcionamento e serão acoplados na versão final do portal.

Google Analytics foi o eleito e acoplado com sucesso, após uma pesquisa exaustiva sobre as metodologias que melhor se adaptassem a estrutura do Portal-Bioinfo. Foi observado que novas funcionalidades como o número e origem de acessos (estatísticas) deveriam ser mostrados no Portal visando uma melhor análise do uso por parte de pesquisadores no Brasil e no exterior. Informações sobre o desempenho computacional e o uso dos recursos do Santos Dumont poderão ser mostrados também na interface.

Foram implementados alguns métodos para fazer as requisições dos números de acessos por Países e Estados, usando assim, os valores retornados para abastecer a API do GeoCharts com dados reais de acessos ao portal.

Foi implementado também a API GeoCharts a algumas páginas específicas do portal, mostrando os acessos (por Países e por estados do Brasil) que o portal tem desde a sua criação até o momento em que a requisição for feita.

Transição das versões PHP5 para PHP7. A atualização da linguagem de programação de PHP5 para PHP7, ainda se encontra em processo de execução. Contudo, algumas atualizações já foram feitas, como: métodos de conexão e consultas ao banco de dados, e algumas outras funções especificas, como funções de percorrer e ordenar vetor, etc.

CONCLUSÕES

Foram detectados pontos críticos como versionamento (ou até falta) de scripts que poderiam ser melhorados e que podem estar relacionados a gargalos no desempenho do Portal-Bioinfo. Concluímos que a atualização nas tecnologias utilizadas pelo portal, como: PHP7 podem tornar o Portal mais rápido, seguro e eficiente. O trabalho em andamento foi apresentado na forma de Pôster na Jornada de Iniciação Científica do LNCC, realizada em fevereiro de 2019.

REFERÊNCIAS BIBLIOGRÁFICAS

Gesing S, Krüger J, Grunzke R, Herres-Pawlis S, Hoffmann A. Using Science Gateways for Bridging the Differences between Research Infrastructures. Journal of Grid Computing. 2016;14:545–57

Gesing S, Nabrzyski J, Jha S. Gateways to high-perfomance and distributed computing resources for global health challenges. In: 2014 IEEE Canada International Humanitarian

Technology Conference - (IHTC). Montreal, QC: IEEE; 2014. p. 1–5. doi:10.1109/IHTC.2014.7147530

Gomes ATA, Bastos BF, Medeiros V, Moreira VM. Experiences of the Brazilian national high-performance computing network on the rapid prototyping of science gateways:

SGW-2013 SPECIAL ISSUE. Concurrency and Computation: Practice and Experience. 2015;27:271–89

Mondelli M, de Souza M, Ocaña K, Vasconcelos A, Gadelha L. HPSW-Prof: A Provenance-Based Framework for Profiling High Performance Scientific Workflows. In: Proceedings of

Satellite Events of the 31st Brazilian Symposium on Databases (SBBD 2016). Bahia, Brazil. p. 117–22