Parallel algorithm Friends-of-Friends to identify galaxies and cluster of galaxies for dark matter halos

Renata S. Rocha Ruiz[1], Haroldo F. Campos Velho[1], Cesar Augusto Caretta[2]

[1]Pós-Graduação em Computação Aplicada – Instituto Nacional de Pesquisas Espaciais (INPE)
São José dos Campos – SP – BRASIL

[2]Departamento de Astronomía – Universidad de Guanajuato
Guanajuato – México

{renata,haroldo}@lac.inpe.br, caretta@astro.ugto.mx

Abstract. *Parallel implementation of a percolation method (Friend-of-Friends) for classifying two astronomical structures: galaxies, and cluster of galaxies for dark matter halos from N-body simulations. From this classification, it is computed the spectra for gravitational potential energy for investigation of the turbulent-like physical dynamics for cosmological evolution. The performance of the parallel version of the algorithm is evaluated by speedup and efficiency from a region of the Virgo project with 318.133 particles.*

Resumo: *Implementação de uma versão paralela do algoritmo de percolação "Friends of Friends" para identificar halos de matéria escura de galáxias e de aglomerados de galáxias em dados provenientes da simulação de N-corpos. A partir desta classificação é feito uma análise do espectro de energia potencial gravitacional destes halos para investigar uma possível dinâmica turbulenta na evolução cosmológica. O desempenho da versão paralela foi avaliado através do cálculo do speedup e da eficiência sobre uma região do projeto Virgo com 318.133 partículas.*

1. Introdução

O entendimento sobre a formação das estruturas presentes no Universo é uma das questões centrais da astrofísica moderna. Em geral, o modelo padrão de formação dessas estruturas é baseado no critério de instabilidade de *Jeans*. Segundo tal critério, perturbações de grande escala tendem a serem cada vez maiores, logo, a densidade na perturbação aumenta continuamente com tempo. Esta instabilidade cria uma concentração de matéria que pode evoluir para formar estrelas, galáxias, aglomerados de galáxias, super aglomerados de galáxias e assim por diante [Madsen 1996, Tatekawa 2005, Hawley e Holcomb 2005].

Uma proposta interessante é a possibilidade de existência de um comportamento similar a turbulência no processo de agrupamento da matéria para formação das grandes estruturas (galáxias, aglomerados, super aglomerados, filamentos, etc.). Essa idéia está associada a uma proposta de Zel'Dovich, em que a evolução cosmológica poderia ser descrita de maneira similar a uma dinâmica turbulenta [Shandarin e ZelDovich 1989].

Caretta et al. (2008) realizaram uma análise preliminar usando halos de matéria escura de galáxias e de aglomerados de galáxias na busca de assinaturas da turbulência na

evolução cósmica do Universo. Nesta análise, foram utilizados dados de simulação de N-corpos provenientes do Consórcio Virgo[1]. A simulação de N-corpos exerce um papel fundamental no estudo da formação da estrutura cósmica. Nesta metodologia, calcula-se a evolução das estruturas de matéria escura a partir de uma época primordial, cuja distribuição de matéria é produzida pela evolução linear de flutuações primordiais. Os resultados obtidos são comparados com dados observacionais [Efastathiou et al. 1985, Bertschinger 1998, Jenkins et al. 1998].

Dentre os quatro modelos de evolução da matéria escura que são simulados pelo projeto Virgo, Caretta et al. 2008 trabalharam com o modelo Λ-CDM (Λ: constante cosmológica; CDM: *Cold Dark Matter*). Os parâmetros numéricos que caracterizam esse modelo são: um cubo com 239.5 h^{-1} Mpc (megaparsecs) de lado e 256^3 partículas com massa individual de $6.86\times10^{10}\ h^{-1}\ M_\Theta$ (h: taxa de expansão; $h=H_0/(100$ Km/s); H_0: constante de Hubble; M_Θ: massa solar $\sim 1.98892\times10^{30}$ Kg).

Os halos de matéria escura foram identificados empregando-se o algoritmo de agrupamento *Friends of Friends* (FoF) [Huchra e Geller 1982, Einasto et al. 1984], também conhecido como técnica de percolação. A partir das propriedades desses halos foi avaliado o espectro de energia potencial gravitacional. Os resultados indicaram que, para haloes de galáxias, este espectro pode ser descrito em termos de uma lei de potência do tipo $-5/3$ em um intervalo de 15 à $50h^{-1}$ Mpc. Isto pode ser um indicativo de um comportamento turbulento, uma vez que a principal característica da turbulência é a existência de uma cascata de energia cinética, cujo comportamento do espectro para o subdomínio inercial é expressa por uma lei de potência de $-5/3$. Esta é conhecida como a lei de Komolgorov de 1941 [Frisch 1995].

Todavia, devido ao elevado custo computacional do algoritmo *Friends of Friends* e dos demais algoritmos necessários ao cálculo do espectro de energia potencial dos grupos identificados, somente um volume pequeno do domínio total de integração do modelo computacional do projeto Virgo foi analisado. Desse modo, neste trabalho apresentamos uma versão paralela do algoritmo FoF que será utilizada para aprofundar os estudos realizados anteriormente, considerando o volume total de integração do projeto Virgo, bem como, dados da Simulação Milênio [Springel et al. 2005] que possui maior resolução.

As demais seções deste trabalho estão divididas da seguinte maneira: Na Seção 2 tem-se a descrição do algoritmo *Friends of Friends*, a Seção 3 apresenta uma breve descrição da estratégia de paralelização utilizada. Uma análise do desempenho do algoritmo paralelo é apresentada na Seção 4. Finalmente a Seção 5 é reservada as considerações finais e os trabalhos futuros.

2. Algoritmo *Friends of Friends*

Um dos métodos mais utilizados na simulação de N-corpos para se determinar estruturas no Universo é o algoritmo de percolação FoF [Huchra e Geller 1982, Caretta et al. 2008]. A idéia básica deste algoritmo é a seguinte: considere uma esfera de raio R ao redor de cada partícula do conjunto total. Se dentro dessa esfera existir outras partículas, elas serão consideradas pertencentes ao mesmo halo e serão chamadas de amigas. Em seguida, toma-se uma esfera ao redor de cada amiga e continua o

[1] http://www.mpa-garching.mpg.de/Virgo/data_download.html

procedimento usando a regra "qualquer amigo de meu amigo é meu amigo". O procedimento pára quando nenhuma amiga nova pode ser adicionada ao grupo. Em outras palavras, o algoritmo FoF agrupa partículas que são separadas por um certo tamanho de ligação *l*. Este tamanho, freqüentemente, é dado por *b* vezes a separação média entre as partículas, os valores de *b* e *l* dependem da natureza da aplicação [Caretta al. 2008]. Os grupos resultantes são limitados por uma superfície de densidade constante de aproximadamente:

$$\frac{n}{\bar{n}} = \frac{2}{(4/3)\pi l^3}\frac{1}{n} = \frac{3}{2\pi l^3}\bar{l}^3 = \frac{3}{2\pi}\frac{1}{b^3} \approx \frac{1}{2b^3} \tag{2.1}$$

em que *n* é o número de objetos e $\bar{l}$ é o número médio de partículas na região considerada. As principais vantagens do FoF são a simplicidade, a reprodutibilidade e a capacidade de detectar *haloes* de qualquer forma [Caretta et al. 2008].

O seguinte pseudocódigo apresenta uma breve descrição do algoritmo FoF.

```
Início do Programa
Leitura dos dados ()
Alocação de Memória ()
L = 1
para i de 1 até # de partículas faça
    igru[i] ← L
    enquanto igru[i] ≠ 0 faça
    i ← i+1
    fim – enquanto
    para j de i  até # de partículas faça
        se igru[j] = L então
        para k de j +1 até # de partículas faça
            se igru[k] = 0 então
            calcula distância de j até k
                se distância < Raio de Percolação então
                igru[k] ← L
                fim – se
            fim – se
        fim –  se
        fim –  para
    fim – para
    L ← L +1
fim – para
Escreve saída ()
```

3. Estratégia utilizada na paralelização

De um modo geral, podemos dizer que a computação paralela é o uso simultâneo de vários processadores para resolver um determinado problema. Os modelos de programação paralela utilizam o conceito de processos, desse modo, o paralelismo é obtido por meio da execução simultânea de um conjunto de processos [Tomita 2004]. Os principais modelos de paralelismo são a troca de mensagens entre processos, cujo

modelo padrão é a biblioteca MPI (Message-Passing Interface) [Pacheco 1997] e o modelo *fork-join* de processos, cujo modelo padrão é o OpenMP [Chapman et al. 2007].

Neste trabalho foi utilizado o modelo de troca de mensagens, ou seja, as atividades dos processadores são coordenadas por meio do envio e do recebimento de mensagens. A MPI fornece funções específicas que permitem essa interação.

A versão paralela do algoritmo *Friends of Friends* implementada neste trabalho consiste na divisão do domínio em vários processadores. O processador mestre faz a leitura dos dados de entrada e o balanceamento de carga, em seguida, usando rotinas da biblioteca MPI, envia os dados correspondentes para cada processador que realiza o agrupamento dos mesmos usando o FoF. Após a identificação dos grupos, cada processador os envia para o processador mestre que realiza o pós-processamento e escreve o arquivo de saída correspondente. A necessidade do pós-processamento se deve ao fato de que partículas pertencentes a um mesmo grupo podem ser enviadas a processadores diferentes, ocasionando a divisão de um grupo em diferentes grupos. Assim, um tratamento é feito nas interfaces para corrigir esse problema, onde as partículas pertencentes a um mesmo grupo, mas que foram enviadas a processadores diferentes são identificadas e associadas aos seus respectivos grupos.

4. Resultados e Análise de desempenho

Geralmente as duas medidas mais utilizadas na análise do desempenho de um programa paralelo na resolução de um determinado problema são o *speedup* e a eficiência [Pacheco 1997]. Basicamente, o *speedup* é a razão entre o tempo de execução gasto no problema serial e o tempo de execução gasto no programa paralelo. Assim, se $T_\sigma(n)$ representa o tempo de execução na solução serial e $T_\pi(n, p)$ o tempo de execução na solução paralela com p processadores, então o *speedup* do programa paralelo é:

$$S(n, p) = \frac{T_\sigma(n)}{T_\pi(n, p)} \tag{4.1}$$

Para um valor fixo de p, usualmente tem-se: $0 < S(n, p) \leq p$. Um *speedup* é chamado linear ou ideal quando $S(n, p) = p$. Se $S(n, p) > p$ o *speedup* é chamado super linear.

Uma alternativa ao *speedup* é a eficiência. De acordo com Pacheco (1997), a eficiência é uma medida da utilização de processadores em um programa paralelo, relativo ao programa serial. Ela é definida conforme equação 4.2:

$$E(n, p) = \frac{S(n, p)}{p} = \frac{T_\sigma(n)}{pT_\pi(n, p)} \tag{4.2}$$

Se $E(n, p) = 1$, significa que o programa paralelo tem uma eficiência de 100%.

Para validar a implementação paralela foi realizado uma aplicação desta versão sobre uma região cúbica do projeto Virgo com 60 $Mpch^{-1}$ de lado e 318.133 partículas no *redshift* $z = 0$. As simulações foram executadas em um cluster HP XC, do Centro de Processamento de Alto Desempenho (C-PAD) do INPE de São José dos Campos (sistema computacional financiado pelo programa Equipamentos Multi-usuários da FAPESP). Esta máquina é baseada em uma arquitetura escalar AMD-Opteron 2.2 GHz

e interconexão de alta velocidade InfiniBand® com uma largura de banda de 2.5 Gbps. A máquina paralela possui 27 nós, dos quais 1 é reservado ao acesso, 1 para o armazenamento de dados e 25 para o processamento, totalizando 112 CPUs (Os nós 1 a 23 têm 4 CPUs, os nós 24 e 25 têm 8 CPUs e os nós 26 e 27 têm 2 CPUs.). O sistema permite a execução simultânea de até 112 jobs gerenciados em fila de execução. O nó de armazenamento tem um arranjo de discos RAID 6 com 4.5 TB de capacidade disponível para dados de usuários. Os códigos computacionais desenvolvidos foram implementados usando a linguagem de programação C e foram executados com compiladores Intel.

Na Tabela 1 pode-se observar o tempo gasto na execução paralela do algoritmo FoF, o tempo gasto na realização do pós-processamento e o tempo total da execução do programa sobre os dados acima citados. Já na Tabela 2 pode-se observar o *speedup* obtido com a execução do FoF em vários processadores, o *speedup* com o tempo total, considerando também o tempo gasto na realização do pós-processamento e finalmente a eficiência obtida com a paralelização. A Figura 1 exibe uma comparação entre o *speedup* ideal e os *speedups* obtidos neste trabalho. Conforme se pode observar o *speedup* obtido com a implementação paralela do algoritmo FoF apresenta um comportamento acima do *speedup* ideal, ou seja, *speedup* super linear. Conseqüentemente se pode observar na tabela 2 que a eficiência desta implementação também apresentou um comportamento acima do ideal. Finalmente, a Figura 2 exibe o comportamento do espectro de energia potencial gravitacional para halos de matéria escura de galáxias, considerando 6 diferentes *redshifts*. A linha sólida nesta figura representa a inclinação -5/3. Pode-se observar nesta figura que as inclinações obtidas para os diferentes instantes da evolução cosmológica são próximas a -5/3.

Tabela 1. Tempo gasto na execução do algoritmo considerando um cubo com 60 Mpch^{-1} de lado e 318.133 partículas.

Nº Processadores	Tempo FoF (s)	Tempo Pós (s)	Tempo Total (s)
Serial	536,867		536,867
2	180,467	10,953	191,42
3	82,403	10,458	92,861
4	48,295	10,641	58,936
7	18,633	16,298	34,931
9	15,242	8,667	23,909
12	10,842	10,046	20,888
19	8,706	10,578	19,284

Tabela 2. *Speedup* e eficiência obtidos variando-se o número de processadores na execução do algoritmo, considerando um cubo com 60 Mpch^{-1} de lado e 318.133 partículas.

Nº Processadores	*Speedup FoF*	*Speedup* Tempo total	Eficiência Tempo total
2	2,975	2,805	1,444
3	6,515	5,781	2,083
4	11,116	9,109	2,634
7	28,813	15,369	3,659
9	35,223	22,454	3,681
12	49,517	25,702	3,831
19	61,666	27,84	3,051

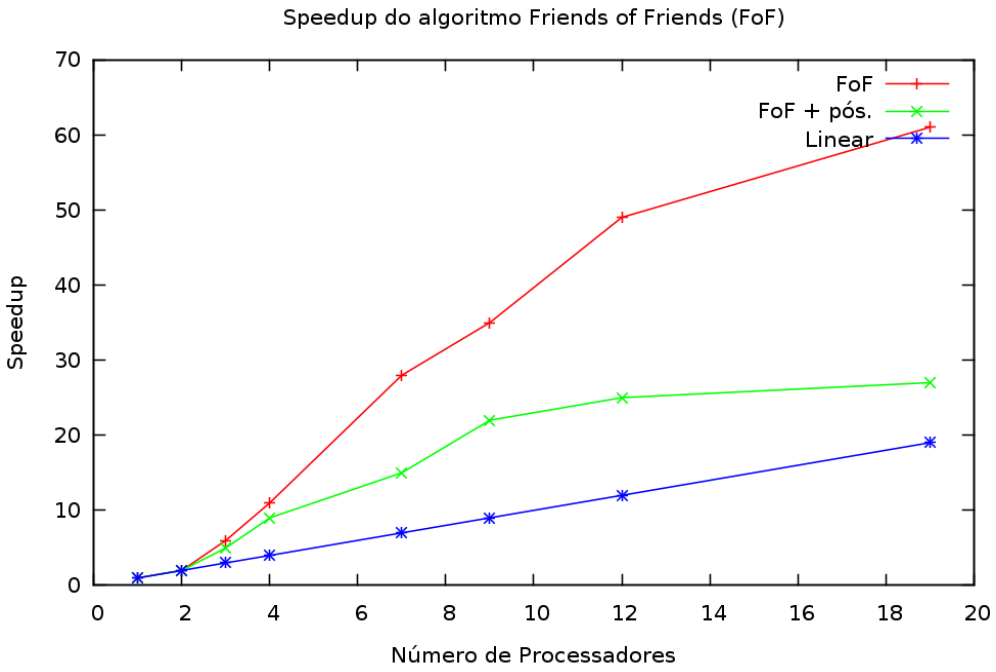

Figura 1. *Speedup* considerando um cubo com 60 Mpch^{-1} de lado e 318.133 partículas. A linha vermelha representa o *speedup* do FoF sem considerar o tempo de pós-processamento. Já a linha verde representa o *speedup* considerando também o tempo gasto no pós-processamento.

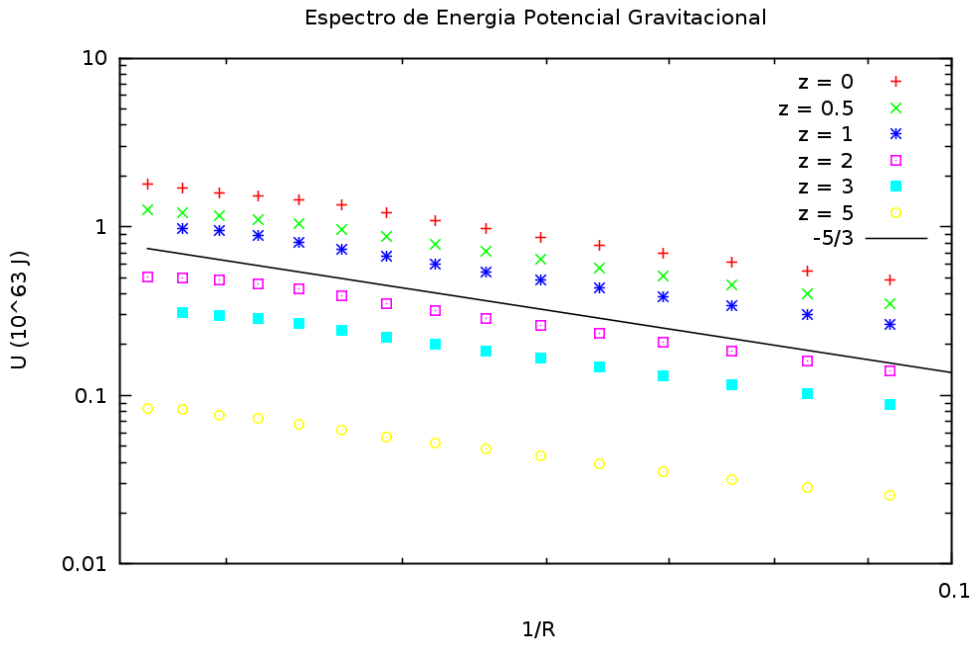

Figura 2. Espectro de energia potencial gravitacional para halos de galáxias considerando um cubo com 60 Mpch^{-1} de lado e 318.133 partículas.

5. Considerações Finais

Neste trabalho foi apresentado um esquema de paralelização do algoritmo FoF cuja aplicação é necessária na identificação de halos de matéria escura para posterior análise do espectro de energia potencial gravitacional em busca de comportamento turbulento no processo de formação das grandes estruturas do Universo. A análise do desempenho via *speedup* e eficiência demonstraram que a versão paralela implementada é plenamente satisfatória, apresentando um comportamento acima do ideal. O espectro potencial gravitacional obtido com o cubo de 60 Mpch^{-1} de lado e 318.133 partículas apresenta um coeficiente de inclinação próximo ao -5/3, isto pode indicar um comportamento turbulento na formação das estruturas do Universo.

Os próximos passos no desenvolvimento desse trabalho são: a implementação paralela das rotinas que calculam as propriedades dos halos identificados pelo FoF, montagem de uma grade computacional nos seguintes departamentos: Laboratório Associado de Computação e Matemática Aplicada (LAC) e Divisão de Astrofísica (DAS), ambos departamentos do INPE e no departamento de Astronomia da Universidade de Guanajuato (México), avaliação do espectro de energia potencial gravitacional para a simulação total do projeto Virgo com o cubo de 239.5 h^{-1} Mpc de lado e 256^3 partículas e para dados da Simulação Milênio.

Agradecimentos. Os autores agradecem ao Consórcio Virgo pelo fornecimento dos dados utilizados e a FAPESP pela bolsa de doutorado processo nº: 2007/54133 – 0.

Referências

Bertschinger, E. (1998). Simulations of Structure Formation in the Universe. Annu. *Rev. Astron. Astrophys*, v. 36, p. 599 -654.

Caretta, C. A., Rosa, R.R, Campos Velho, H. F., Ramos, F. e Makler, M. (2008). Evidence of Turbulence-like universality on the formation of galaxy-sized dark matter halos. *Astronomy & Astrophysics,* v. 487, p. 445 – 451.

Chapman, B., Jost, G. and van der Pas, R. (2007). *Using OpenMP: Portable Shared Memory Parallel Programming*. EUA: The MIT Press.

Efastathiou, G., Davis, M., White, S. D. M., Frenk, C.S. (1985). Numerical techniques for large cosmological N-corpos simulations. *Astrophysical Journal Supplement Series*, v. 57, p. 241-260.

Einasto, J. , Klypin, A. A., Saar, E., Shandarin, S. F. (1984). Structure of Superclusters and Supercluster Formation – III. Quantitative Study of the Local Supercluster. *Mon. Not. R. astr. Soc.*, v. 206, p. 529 – 558.

Frisch, U. (1995). *Turbulence: The Legacy of A.N. Kolmogorov*. New York: Cambridge University Press.

Hawley, J. F., Holcomb, K. A. (2005) *Foundations of Modern Cosmology*. 2 ed. Oxford New York: Oxford University Press.

Huchra, J. P. e Geller, M., J. (1982). Groups of Galaxies I. Nearby groups . *The astrophysical*, v. 257, p. 423 – 437.

Jenkins, A., Frenk, C. S., Pearce, F. R., et al. (1998). Evolution of Structure in Cold Dark Matter Universe. *The Astrophysical Journal*, v. 499, p. 20 – 40.

Madsen, M., S. (1996). *The Dynamic Cosmos – Exploring the Physical Evolution of the Universe*. 1. ed. New York, NY, USA: Chapman & Hall, 144 p.

Pacheco, P. S. (1997). *Parallel Programming with MPI*. San Francisco: Morgan Kaufmann Publishers, Inc.

Shandarin, S. F., Zeldovich, Y. (1989) The Large-Scale of the Universe: Turbulence, Intermittency, Structure in a Self-gravitating Medium. *Reviews of Modern Physics*, v. 61, p. 185 - 220.

Springel, V., Simon, D. W., Jenkins, A., Frenk, C. S., Yoshida, N., Gao, L., Navarro, J., Thacker, R., Croton, D., Helly, J., Peacock, J. A., Cole, S., Thomas , P., Couchman, H., Evrard, A., Colbergm, J., Pearce, F.(2005). Simulating of the formation, evolution and clustering of galaxies and quasars. *Nature*, v. 435, p. 629 - 636.

Tatekawa, T. (2005). Langrangian Perturbations theory in Newtonian Cosmology. *ArXiv:astro-ph/0412025v4*.

Tomita, S. S. (2004). *Metodologia Para Paralelização de Programas Científicos*. Dissertação(Mestrado em Computação Aplicada), 96 p. São José dos Campos: INPE.