Paralelização do Modelo de Cinética Química Atmosférica do CPTEC/INPE Para Utilizar Placas Gráficas

Alex de A. Fernandes[1], Stephan Stephany[2], Jairo Panetta[3]

[1]Programa de Mestrado em Computação Aplicada – CAP
Instituto Nacional de Pesquisas Espaciais – INPE

[2]Laboratório Associado de Computação e Matemática Aplicada – LAC
Instituto Nacional de Pesquisas Espaciais – INPE

[3]Centro de Previsão de Tempo e Estudos Climáticos- CPTEC
Instituto Nacional de Pesquisas Espaciais - INPE

alex.fernandes@cptec.inpe.br, stephan@lac.inpe.br,
jairo.panetta@cptec.inpe.br

Abstract. *The changes in the chemical composition of the atmosphere due to reactions between its components have significant impact on the amount of radiation absorbed by the atmosphere. A chemical kinetics model can be coupled to a numerical model of weather or climate, but their computational cost is very high, hindering their operational use. This paper investigates the parallelization of model chemical kinetics CPTEC/INPE using graphic accelerator cards.*

Resumo. *A alteração da composição química da atmosfera devida às reações entre seus componentes tem impacto significativo na quantidade de radiação absorvida pela atmosfera. Um modelo de cinética química pode ser acoplado a um modelo numérico de previsão de tempo ou climático, mas seu custo computacional é muito alto, dificultando seu uso operacional. Este trabalho investiga a paralelização do modelo de cinética química do CPTEC/INPE utilizando placas aceleradoras gráficas. Para tanto, pretende-se avaliar as interfaces de programação CUDA e OpenACC para obtenção de desempenho compatível com o uso operacional do modelo.*

Palavras-chave: *placas aceleradoras gráficas, modelo de cinética química atmosférica, modelo numérico de previsão de tempo, monitoramento ambiental.*

1. Introdução

Novos elementos químicos são lançados na atmosfera a cada instante, alterando a composição da mesma. Queimadas, emissões de veículos movidos à combustão e gases provenientes de rebanhos de animais, são alguns exemplos de processos que geram

novos elementos. Uma vez depositados na atmosfera, estes elementos são transportados pela atmosfera por vários quilômetros, podendo uma queimada em determinada área rural, afetar de forma impactante uma grande metrópole. Os elementos químicos da atmosfera também sofrem reações químicas, gerando novos elementos na atmosfera. Estas reações normalmente ocorrem por fotólise (reação pela radiação solar) ou cinética. Neste contexto, o modelo de química do Centro de Previsão de Tempo e Estudos Climáticos (CPTEC) do Instituto Nacional de Pesquisas Espaciais (INPE), atua acoplado a um modelo atmosférico, calculando as reações na atmosfera de acordo com mecanismos químicos que delimitam quais reações podem ocorrer, os elementos participantes destas reações e os possíveis elementos gerados, determinando um estado futuro da atmosfera. Um modelo de transporte, o qual não será tratado neste trabalho, faz o transporte dos poluentes pela grade do modelo atmosférico.

Conhecer a composição química da atmosfera é de suma importância para a população, visto que afeta diretamente a saúde e a economia.

Poluentes atingem as porções mais profundas do sistema respiratório e atravessam a barreira epitelial, desencadeando processos inflamatórios [Arbex et al 2004]. Elementos oxidantes, como o ozônio troposférico, além de afetar a saúde da população também afetam a economia, atrapalhando o desenvolvimento de plantações, e deteriorando com o passar do tempo, materiais como o concreto e a borracha.

Os poluentes da atmosfera também impactam no balanço energético do planeta, ao influenciarem na absorção de radiação e influenciando no ciclo hidrológico do planeta. Neste ponto é interessante a integração/acoplamento entre o modelo de química e o modelo atmosférico.

No entanto, a execução do modelo de cinética química demanda um processamento adicional, o que torna seu uso operacional inviável. Dessa forma o uso de GPUs atenuaria essa sobrecarga de processamento, permitindo a execução do modelo CCATT-BRAMS mesmo com grades mais refinadas. Pretende-se investigar o uso das interfaces de programação CUDA e OpenACC, dois paradigmas de programação aplicáveis às placas gráficas.

2. Modelo de Cinética Química

A influência dos poluentes na atmosfera é modelada por meio do acoplamento a um modelo numérico de previsão de tempo ou climático de modelos de transporte, de emissão e de cinética química, os quais se aplicam aos poluentes. Neste trabalho aborda-se especificamente um modelo de cinética química atmosférica, o modelo do CPTEC.

No CPTEC, encontra-se sendo executado operacionalmente, o sistema de monitoramento ambiental CCATT (Coupled-Chemical, Aerosol and Tracer Transport) [Hoelzemann 2008], o que acopla o modelo numérico de previsão de tempo de mesoscala BRAMS (Brazilian Developments on the Regional Atmospheric Modeling System) [Freitas et al 2007] ao CATT (coupled Aerosol and Tracer Transport Model), a um modelo de emissões e ao modelo de cinética química do CPTEC. Pretende-se desta forma, prever a dinâmica química dos poluentes presentes na atmosfera, disponibilizando previsões sobre a qualidade do ar e estimando de uma melhor forma o balanço energético da atmosfera. O CCATT é um modelo executado em tempo real, tridimensional e Euleriano, que prognostica em tempo real a concentração de contaminantes atmosféricos de forma simultânea e totalmente consistente com o estado

atmosférico simulado pelo BRAMS [Freitas et al, 2009].

$$\underbrace{\frac{\partial \bar{s}}{\partial t} + \bar{u}\frac{\partial \bar{s}}{\partial x} + \bar{v}\frac{\partial \bar{s}}{\partial y} + \bar{w}\frac{\partial \bar{s}}{\partial w}}_{I} = \underbrace{-\frac{1}{\rho_0}\left(\frac{\partial \rho_0 \overline{u's'}}{\partial x} + \frac{\partial \rho_0 \overline{v's'}}{\partial y} + \frac{\partial \rho_0 \overline{ws'}}{\partial z}\right)}_{II} + \underbrace{\overline{Q_s}}_{III} \qquad (1)$$

No caso, a equação da continuidade (Eq. 1), que expressa conservação da massa dos 47 componentes químicos atmosféricos considerados, deve ser resolvida a cada *timestep* (passo de tempo) para cada ponto da grade. Nessa equação, a parte relativa à cinética química costuma tomar de 50% a 95% do tempo de processamento [Zangh et al 2011; Linford et al 2009].

A partir da equação 1, para cada ponto de grade, é gerada uma matriz com as concentrações de cada elemento, que por sua vez são definidos pelo mecanismo químico (no CPTEC 47 elementos estão presentes no mecanismo químico operacional). Estas concentrações possuem ordens de grandeza muito discrepantes, gerando um sistema de equações do tipo *stiff* (rígido). Para a solução deste sistema é utilizado o método de Rosenbrock. Interessante citar que no modelo de química cada ponto de grade, ou sistema a ser resolvido, não tem relação com os pontos vizinhos, sendo as concentrações para o próximo *timestep* dependentes apenas de sua concentração local.

Considerando-se que o modelo operacional CCATT-BRAMS do CPTEC/INPE possui uma grade de 340x370x42 pontos (um total de 5.283.600 pontos), e um *timestep* de 160 segundos e que ocorrem 540 execuções para cada dia de previsão, chega-se a conclusão que 5.283.600 sistemas devem ser resolvidos 540 vezes para que se tenha apenas um dia de previsão. Imaginando-se que as grades dos modelos de previsão numérica de tempo tendem a serem cada vez mais refinadas, e consequentemente os *timesteps* diminuídos, pode-se dizer que futuramente e com as técnicas hoje utilizadas, será inviável executar o modelo de química do CPTEC/INPE operacionalmente, isto sem levar em consideração que o método de Rosenbrock utilizado, resolve o sistema em, no mínimo, 4 ciclos de iterações completas.

Tabela 1. Relação entre a resolução do modelo CCATT-BRAMS, passo de tempo e número de pontos de grade.

Resolução	NX	NY	NZ	Timestep (s)	Nº Pontos	Execuções para 1 dia de previsão
20 Km	340	370	42	160	5.283.600	540
10 Km	680	740	42	80	21.134.400	1080
05 Km	1360	1480	42	48	84.537.600	1800

Tabela 2. Relação entre os tempos de execução do modelo de cinética química e o modelo operacional CATT-BRAMS (20 km) por timestep.

Timestep	Tempos do Módulo de Cinética Química (s)	Tempo Total (s)	%
1	663,78	744,88	89,11
2	635,74	717,12	88,65
3	631,81	716,41	88,19

3. Paralelização Usando GPUs (*Graphics Processing Units)*

Desde 2003 as GPUs têm aumentado a capacidade de operações em ponto flutuante a taxas muito superiores quando comparados aos CPU multicore [Kirk; Hwu 2011], conforme pode ser visto na figura 1.

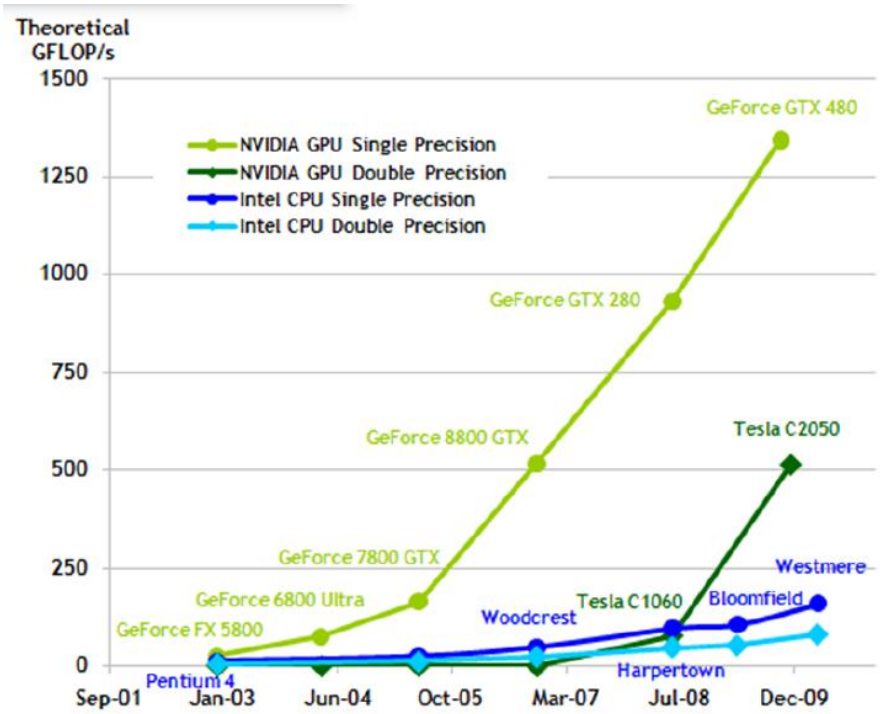

Figura 1. Evolução da capacidade do pico de processamento das GPUs vs CPUs.

A razão para a discrepância da capacidade de operações de uma CPU e de uma GPU, é a quantidade de unidades especificas para processamento presentes em um único chip, ocupando uma área muito maior proporcionalmente do que as unidades de controle de fluxo e memórias cache [CUDA C Programming Guide 2011].

Devido à grande capacidade de processamento das GPUs e à natureza paralela da mesma (com o processamento de múltiplos *threads*), o mercado de HPC (High

Performance Computing) começou a migrar áreas críticas de processamento de determinados programas para o uso com GPU, melhorando drasticamente o desempenho de programas e levando a indústria a criar novas tecnologias e facilidades para a programação em GPU, como por exemplo, as interfaces/extensões de programação CUDA, OpenGL e OPENACC.

Na busca de uma solução para a execução do modelo de química do CPTEC/INPE, o uso de GPUs torna-se promissor, visto que o paralelismo é grande já que não há dependências de dados entre os pontos.

As GPUs escolhidas para uso neste trabalho são de fabricação da Nvidia, assim como as interfaces de programação escolhidas foram CUDA e OPENACC.

Basicamente, as GPUs Nvidia são construídas agrupando múltiplos *cores* (núcleos) em *Stream Multiprocessors* (SM) que executam instruções baseadas no modelo *Single Instruction Multiple Threads* (SIMT). Sendo assim, a cada ciclo, uma mesma instrução é executada concorrentemente para diferentes dados pelos múltiplos *threads*. Para cada GPU, existe um conjunto mínimo de *threads* que pode ser processado em esquema SIMT, denominado *warp*. Comumente são 32 *threads* por *warp* e o número total de *threads* é dividido em *warps* e escalonado para execução nos SMs. Cada *warp* é executado simultaneamente por um SM. Por exemplo, um bloco com um total de 160 *threads* alocado a um SM é dividido em 5 *warps* que são então executados nesse SM em sequência.

Todos os SM contam, além dos núcleos, com unidades de controle de *threads*, unidades para cálculo de funções transcendentais, unidades de *load/store* e ainda com memória própria. Já na placa propriamente dita, ainda existem distribuidores de *threads*, memória compartilhada, entre outros.

Apesar da capacidade de processamento das GPUs ser maior que as CPUs, dois fatores principais devem ser levados em consideração em respeito à programação nestes dispositivos:

- Existe um gargalo na transferência dos dados da CPU para a GPU, o que pode inviabilizar o uso das GPUs caso muitos dados tenham de ser migrados.
- Apesar de ser mais rápida do que seria uma memória normal, a memória global das GPUs são acessadas por todos os núcleos da GPU (geralmente centenas), logo a banda de acesso pode ser dividida, no pior caso, entre todos os núcleos se o uso da memória for intensivo, sendo assim o uso desta memória deve ser feito com cautela.

Tendo conhecimento básico das GPUs, foram escolhidas as interfaces de programação CUDA e OPENACC. CUDA é uma interface proprietária da Nvidia, logo funciona apenas em placas do fabricante citado. CUDA incorpora funções em linguagem C (ou Fortran se utilizado o Fortran da Portland group) e estende a linguagem para execução de trechos de programas nas GPUs. Esta extensão é de implementação trabalhosa, devendo se conhecer muito bem o código a ser modificado e otimizado para que se obtenha desempenho.

Já o OPENACC, é um padrão novo proposto pela Cray, CAPS, Nvidia e Portland, inspirado no padrão OpenMP, que utiliza diretivas de paralelização para execução com múltiplos *threads*, e consequentemente permite programar num nível de abstração mais elevado, o que facilita muito o trabalho de portar um código antigo para o uso com aceleradores. Teoricamente, OPENACC é compatível com qualquer placa aceleradora,

desde que o fabricante forneça suporte. O OPENACC deve prover uma portabilidade que facilitará a paralelização de códigos existentes, além de permitir o uso de placas gráficas de qualquer fabricante e não só da NVidia, como no caso do CUDA.

4. Resultados/Considerações

Atualmente, o trabalho proposto encontra-se na fase de avaliação dos códigos do modelo de cinética química, em particular do solver utilizado para a resolução dos sistemas lineares associados, de forma a quantificar melhor o custo computacional desse modelo e do solver.

Há duas possibilidades a serem estudadas para possível implementação:

- Mapear a resolução do sistema linear associado ao modelo de cinética química em cada ponto de grade para cada *thread* a ser executado na GPU. Entretanto, isso implicaria na execução de um número muito grande de *threads* a cada *timestep*. Neste caso, deve-se avaliar o uso da memória na GPU e também da memória principal, uma vez que há um sistema para cada ponto de grade e há muitos pontos de grade.
- Resolver o sistema linear associado ao modelo de cinética química para um conjunto de pontos de grade cada vez, de forma que cada execução do *solver* seria paralelizado por meio de *threads* a serem atribuídos a um SM da GPU. Dessa forma, por exemplo, numa placa gráfica com 16 SMs, como as Fermi, seria possível executar esse modelo para 16 pontos de grade concorrentemente de cada vez.

A escolha de uma destas opções depende das limitações de memória local da placa gráfica e dos SM, da latência relativa às transferências de dados entre a memória principal da máquina e da memória local da placa gráfica, e vice-versa. Essas restrições somente poderão ser avaliadas melhor após a implementação computacional relativa ao uso de GPU.

No tocante a resolução de sistemas esparsos em geral, existe a possibilidade do uso de bibliotecas específicas para GPUs. Este é o caso da biblioteca para resolução de sistemas esparsos *cuSparse*, cujo uso será analisado quanto a sua aplicabilidade para sistemas do tipo *sitff*, que é o caso do modelo de cinética química. Outro importante fato a ser citado, é que como em qualquer paralelização, a sobreposição de processamento e comunicação é desejável, sendo esta uma maneira de atenuar os custos de comunicação relativos às transferências entre as memórias principal e a memória da GPUs e vice-versa.

5. Conclusões

O modelo de cinética química, cuja paralelização é objeto de estudo, é empregado no sistema de monitoramento ambiental CCATT-BRAMS, o qual acopla também o modelo BRAMS, a um modelo de emissões e ao modelo de transporte CATT. Uma vez que o modelo de cinética química constitui um gargalo em termos de tempo de processamento, torna-se necessária sua paralelização, no caso, para execução numa placa aceleradora gráfica e utilizando a interface de programação OpenACC ou CUDA, visando também a comparação de desempenho entre as duas tecnologias.

O uso de computação híbrida ou heterogênea, na qual a CPU é utilizada conjuntamente com aceleradores de processamento, é uma tendência atual, inclusive em

supercomputadores, sendo o uso de placas gráficas com este fim amplamente difundido. O OpenACC, que é uma interface de programação inspirada no OpenMP, ainda está em desenvolvimento, embora se espere que vá se tornar um padrão num futuro próximo, com a portabilidade esperada.

O uso de GPUs pode ser promissor para o problema proposto, trazendo a possibilidade do refinamento das resoluções dos modelos atmosféricos sem que haja degradação do desempenho computacional da previsão ambiental devido ao modelo de cinética química.

Referências

Arbex, M. A.; Cançado, J. E. D.; Pereira, L. A. A.; Braga, A. L. F.; Saldiva, P. H. N. (2004); Queima de biomassa e efeitos sobre a saúde, J Bras Pneumol 2004; 30(2) 158-175 http://www.scielo.br/pdf/jbpneu/v30n2/v30n2a15.pdf Acesso: 20/03/2012

Hoelzemann, J. J.; Longo, K. M.; Elbern, H.; Freitas, S. R.; (2008) An Aerosol And Trace Gas Data Assimilation System For The CCATT-BRAMS Atmospheric Model With Focus On Brazilian Fire Emissions, Conferência Científica Internacional "Amazônia em Perspectiva: Ciência Integrada para um Futuro Sustentável" nos dias 17 a 20 de novembro de 2008, em Manaus.

Freitas S. R.; Longo, K. M.; Silva Dias, M.; Chatfield, R.; Silva DIAS, P.; Artaxo, P.; Andreae M. O.; Grell, G.; Rodrigues, L. F.; Fazenda, A.; Panetta J. (2007) The Coupled Aerosol and Tracer Transport model to the Brazilian developments on the Regional Atmospheric Modeling System, Atmos. Chem. Phys. Discuss., 7, 8525–8569, 2007 www.atmos-chem-phys-discuss.net/7/8525/2007/ , Acesso: 20/03/2012

Zhang, H.; Linford, J. C.; Sandu, A.; Sander, R. (2011) Chemical Mechanism Solvers in Air Quality Models, Atmosphere 2011, 2, 510-532; doi:10.3390/atmos2030510 http://www.mdpi.com/2073-4433/2/3/510/s1 Acesso: 20/03/2012

Linford, J. C.; Vachharajani, M.; Michalakes, J.; Sandu, A. (2009) Multi-core Acceleration of Chemical Kinetics for Simulation and Prediction, Super Computing 2009, http://people.cs.vt.edu/~jLINFORD/?download=sc09-manuscript.pdf Acesso: 20/03/2012

Freitas, S. R.; Longo, K.; Rodrigues, L. F. (2009), Modelagem Numérica da Compsição Química da Atmosfera e seus Impactos no Tempo, Clima e Qualidade do Ar, Revista Brasileira de Meteorologia, v.24, n.2, 188-207, 2009

Kirk, D. B.; Hwu, W. W. (2011); Programming Massively Parallel Processors: A Hands on Approach, ISBN: 978-0-12-381472-2 2011

NVIDIA: NVidia CUDA C Programing Guide v4.1 (2011), http://developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/CUDA_C_Programming_Guide.pdf , Acesso: 20/03/2012