Susceptibilidade a falhas críticas em algoritmos de classificação

Caio Brigagão Lunardi & Paolo Rech (Orientador)
Universidade Federal do Rio Grande do Sul
cblunardi@inf.ufrgs.br

Introdução

Alguns dos algoritmos mais relevantes no cenário da computação de alto desempenho são os de classificação, empregados em aplicações complexas tais quais rastreamento de estrelas, codificação de vídeo, sistemas de segurança, etc. Estas aplicações se baseam, muitas vezes, em implementações baseadas em processadores paralelos, as GPGPUS (Placas de Vídeo de Propósito Geral), possibilitando uma quantia expressiva de cálculos com um baixo consumo de potência.

Estas implementações de algoritmos de classificação em processadores paralelos são, porém, susceptíveis a falhas de dois tipos induzidas por radiação (nêutrons): corrupção silenciosa de dados (SDC) ou interrupção da aplicação (Crash).

Objetivos e Metodologia

O objetivo do trabalho é determinar o impacto e a criticalidade das diversas falhas ocassionadas pela radiação atmosférica em aplicações de classificação de alta performance para dispositivos paralelos.

Foram conduzidos testes utilizando placas NVIDIA Tesla K40c sob o acelerador de partículas disponível no Los Alamos Neutron Science Center (LANSCE). O fluxo de nêutrons disponível simulava as partículas presentes na atmosfera entre 10 e 750 MeV, com 8 ordens de magnitude superior ao presente no nível do mar, conforme a Fig.1.

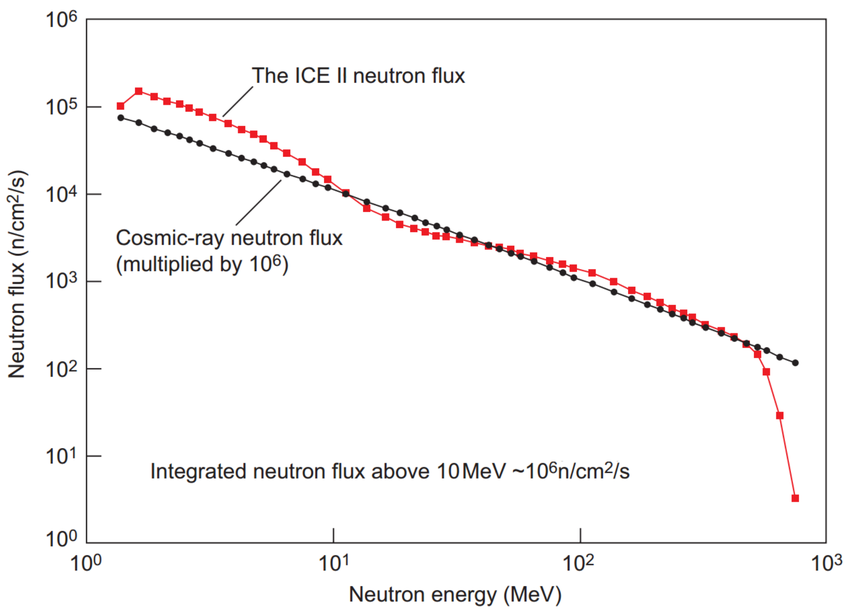

Figure 1: Fluxo de nêutrons no LANSCE comparado ao fluxo de nêutrons ao nível do mar multiplicado por 10^8

Como somente o chipset das GPGPUs foram irradiados, conforme mostra a Fig.2, as memórias não foram afetadas. Para realizar a avaliação da tolerância à falhas de cada aplicativo, foi usada a métrica FIT (Failure in Time).

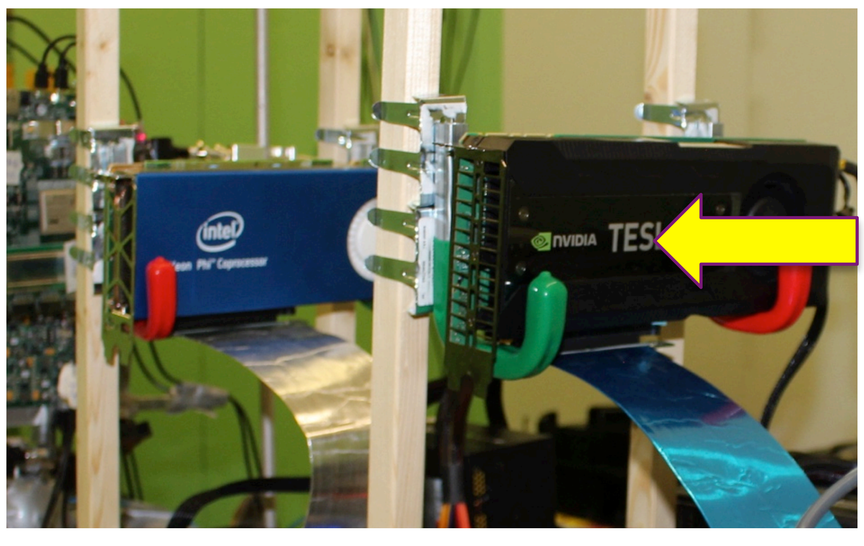

Figure 2: Posicionamento das placas Tesla K40 sob o uxo de nêutrons

Foram empregados três implementações de algoritmos de classificação paralelos, todos em linguagem CUDA: QuickSort, MergeSort e RadixSort. Cada implementação apresenta características particulares, possibilitando uma análise em detalhe das diversas técnicas de otimizações e emprego de recursos disponíveis na GPGPU, tais quais o CDP (CUDA Dynamic Parallelism) e a biblioteca de alto desempenho *Thrust*.

Foram preparados três *benchmarks* com tamanhos de entradas distintos para cada aplicação, de 32, 64 e 128 milhões de elementos inteiros. Isto possibilitou verificar como o utilizo dos recursos da GPGPU impactava na sensibilidade à falhas induzidas por radiação.

Resultados

Os resultados experimentais para o tamanho 64M estão apresentados na Fig.3. De forma geral, pode ser notado que cada código apresenta maior ou menor sensibilidade dependendo do rol de otimizações e técnicas de implementação empregadas.

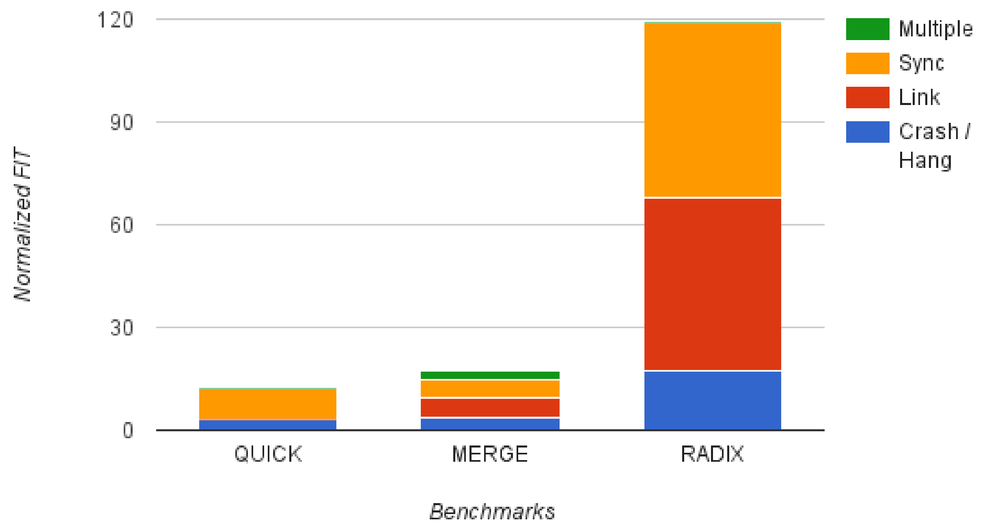

Figure 3: MWBF do código GEMM com diversos tamanhos de input

Conclusão

Foi constatada uma grande variação no tempo de execução empregado pelos três benchmarks para realizar a classificação das mesmas entradas e também na sensibilidade intrínseca de cada implementação.

O RadixSort mostrou ser o mais rápido, porém o mais sensível, enquanto o QuickSort mostrou ser lento mas robusto. Foi possível notar que o MergeSort mostrou uma excelente razão entre ambos fatores.

A análise de criticalidade dos erros foi fundamental na identificação do comportamento das falhas induzidas pela radiação, e serão fundamentais na investigação de técnicas para tratamento destas.

A meta para as próximas pesquisas será a de propor métodos eficazes no tratamento das falhas observadas nesta pesquisa e que possam ser empregadas em um amplo rol de implementações de algorítmos de classificação paralelos.

References

[1] C. Lunardi, H. Quinn, L. Monroe, D. Oliveira, P. Navaux, and P. Rech. On the Error Criticality of Sorting Algorithms for HPC and Large Servers Applications. In *RADECS*, 2016.