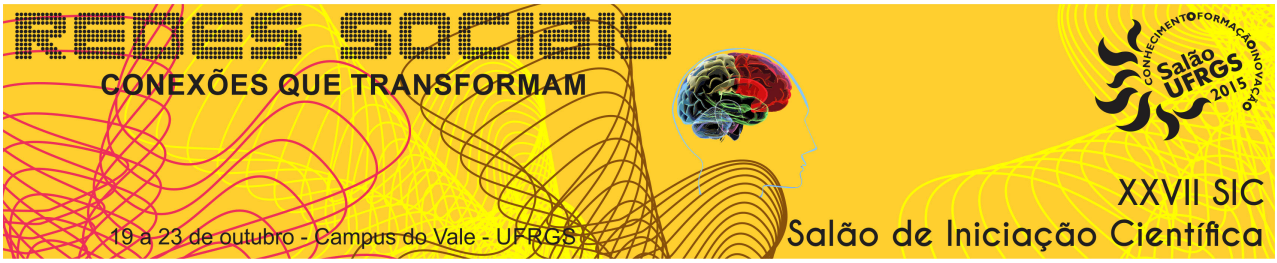


| | |
|---|---|
| **Evento** | Salão UFRGS 2015: SIC - XXVII SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS |
| **Ano** | 2015 |
| **Local** | Porto Alegre - RS |
| **Título** | Tolerância a Falhas em Aplicações Paralelas de Alto Desempenho |
| **Autor** | CAIO BRIGAGÃO LUNARDI |
| **Orientador** | PAOLO RECH |

# Tolerância a Falhas em Aplicações Paralelas de Alto Desempenho

Caio Brigagão Lunardi (Orientador: Paolo Rech)

Universidade Federal do Rio Grande do Sul

Aplicações Paralelas de Alto Desempenho são de extrema relevância para diversos setores da indústria e em âmbitos científicos. Estas aplicações são utilizadas nas simulações dos mais diversos tipos, tais quais simulações físicas de fluidos, colisões, processos em cadeia, estatísticos e modelos de alta precisão, e são utilizadas em super computadores tais quais o Titan, construído no Oak Ridge National Laboratory, Estados Unidos, composto de milhares de GPGPUs (Placas de Vídeo de Propósito Geral), processadores altamente paralelos que podem executar milhares de processos ao mesmo tempo.

A construção de aplicações paralelas de alto desempenho requer uma etapa de busca de otimizações, otimizações estas que aumentam a eficiência, reduzem o consumo de energia e diminuem os tempos de processamento de dados de grande escala. Porém não se sabe muito ao certo sobre a tolerância à falhas induzidas por radiação destas otimizações realizadas em códigos. Em um único processador, estas falhas ocorrem raramente podem ser ignoradas devido à pouca importância de um único erro numa aplicação não crítica. Porém, no super computador Titan, ocorre um erro não detectável no output a casa 8h e um crash devido à radiação a cada 44h, o que é inaceitável. A tolerância à falhas também é fundamental para aplicações críticas, tais quais as empregadas no setor automobilístico e aerospacial.

O trabalho realizado consistiu na experimentação sobre como ocorrem estas falhas em diversos tipos de aplicações paralelas de alta performance, qual sua correlação com as diversas técnicas de otimização empregadas e a potencial resistência a falhas dos diversos componentes da arquitetura das GPGPUs tais quais as utilizadas no Titan. As aplicações utilizadas foram diversos benchmarks tais quais FFT (Fast Fourier Transform), GEMM (Multiplicação de Matrizes), Hotspot (Simulação de distribuição de calor em uma superfície), lavaMD (Simulação de colisão entre blocos) e Kmeans (Algoritmo de agrupamento). Estes benchmarks foram escolhidos em base aos recursos que utilizavam do processador paralelo, para assim poder extender a abrangência deste grupo restrito de aplicações testadas para um âmbito muito maior de classes de aplicações diferentes presentes no mundo real.

Para os testes, foi utilizada a NVIDIA Tesla K20, uma placa de alto desempenho com 2496 unidades de processamento paralelas. Para obter dados reais, foram conduzidos testes sob o acelerador de partículas disponível no Los Alamos Neutron Science Center. O fluxo de nêutrons disponível simulava as partículas presentes na atmosfera entre 10 e 750 MeV, com 10 ordens de magnitude superior ao presente no nível do mar. No caso da GEMM, o tempo médio entre duas falhas induzidas pela radiação aumentou em cerca de uma ordem de magnitude. É notado que as caches são o alvo de boa parte das otimizações, já que seu bom emprego está relacionado à redução no tempo de execução das aplicações, e que sua sensibilidade à falhas provocadas por radiação é acentuada. Porém, como o tempo de execução diminuiu consideravelmente, as otimizações provaram ser um bom método para se aumentar a tolerância a falhas nas aplicações de alta performance, já que muito mais trabalho útil pode ser realizado durante o tempo entre duas falhas.