Hardening Strategies for HPC Applications

Daniel Oliveira[1], Paolo Rech[1], Philippe Olivier Alexandre Navaux[1]

[1]Universidade Federal do Rio Grande do Sul

Abstract. HPC devices reliability is one of the major concerns for supercomputers today and for the next generation. In fact, the high number of devices in large data centers makes the probability of having at least a device corrupted to be very high. In this work, we first evaluate the problem by performing radiation experiments. The data from the experiments give us realistic error rate of HPC devices. Moreover, we evaluate a representative set of algorithms deriving general insights of parallel algorithms and programming approaches reliability. To understand better the problem, we propose a novel methodology to go beyond the quantification of the problem. We qualify the error by evaluating the criticality of each corrupted execution through a dedicated set of metrics. We show that, as long as imprecise computing is concerned, the simple mismatch detection is not sufficient to evaluate and compare the radiation sensitivity of HPC devices and algorithms. Our analysis quantifies and qualifies radiation effects on applications output correlating the number of corrupted elements with their spatial locality. We also provide the mean relative error (dataset-wise) to evaluate radiation-induced error magnitude. Furthermore, we designed a homemade fault-injector, CAROL-FI, to understand further the problem by collecting information using fault injection campaigns that is not possible through radiation experiments. We inject different fault models to analyze the sensitivity of given applications. We show that portions of applications can be graded by different criticalities. Mitigation techniques can then be relaxed or hardened based on the criticality of the particular portions. This work also evaluates the reliability behaviors of six different architectures, ranging from HPC devices to embedded ones, with the aim to isolate code- and architecture-dependent behaviors. For this evaluation, we present and discuss radiation experiments that cover a total of more than 352,000 years of natural exposure and fault-injection analysis based on a total of more than 120,000 injections. Finally, Error-Correcting Code, Algorithm-Based Fault Tolerance, and Duplication With Comparison hardening strategies are presented and evaluated on HPC devices through radiation experiments. We present and compare both the reliability improvement and imposed overhead of the selected hardening solutions. Then, we propose and analyze the impact of selective hardening for HPC algorithms. We perform fault-injection campaigns to identify the most critical source code variables and present how to select the best candidates to maximize the reliability/overhead ratio.

Resumo. A confiabilidade de dispositivos de Processamentos de Alto Desempenho (PAD) é uma das principais preocupações dos supercomputadores hoje e para a próxima geração. De fato, o alto número de dispositivos

em grandes centros de dados faz com que a probabilidade de ter pelo menos um dispositivo corrompido seja muito alta. Neste trabalho,primeiro avaliamos o problema realizando experimentos de radiação. Os dados dos experimentos nos dão uma taxa de erro realista de dispositivos PAD. Além disso, avaliamos um conjunto representativo de algoritmos que derivam entendimentos gerais de algoritmos paralelos e a confiabilidade de abordagens de programação.Para entender melhor o problema, propomos uma nova metodologia para ir além da quantificação do problema. Qualificamos o erro avaliando a importância de cada execução corrompida por meio de um conjunto dedicado de métricas. Mostramos que em relação a computação imprecisa, a simples detecção de incompatibilidade não é suficiente para ava-liar e comparar a sensibilidade à radiação de dispositivos e algoritmos PAD. Nossa análise quantifica e qualifica os efeitos da radiação na saída das aplicações, correlacionando o número de elementos corrompidos com sua localidade espacial. Também fornecemos o erro relativo médio (em nível do conjunto de dados) para avaliar a magnitude do erro induzido pela radiação.Além disso, desenvolvemos um injetor de falhas, CAROL-FI, para entender melhor o problema coletando informações usando campanhas de injeção de falhas, o que não é possível através de experimentos de radiação. Injetamos diferentes modelos de falha para analisar a sensitividade de determinadas aplicações. Mostramos que partes de aplicações podem ser classificadas com diferentes criticalidades. As técnicas de mitigação podem então ser relaxadas ou enrobustecidas com base na criticalidade de partes específicas da aplicação.Este trabalho também avalia a confiabilidade de seis arquiteturas diferentes, variando de dispositivos PAD a embarcados, com o objetivo de isolar comportamentos dependentes de código e arquitetura. Para esta avaliação, apresentamos e discutimos experimentos de radiação que abrangem um total de mais de 352.000 anos de exposição natural e análise de injeção de falhas com base em um total de mais de 120.000 injeções.Por fim, as estratégias de ECC, ABFT e de duplicação com comparação são apresentadas e avaliadas em dispositivos PAD por meio de experimentos de radiação. Apresentamos e comparamos a melhoria da confiabilidade e a sobrecarga imposta das soluções de enrobustecimento selecionadas. Em seguida, propomos e analisamos o impacto do enrobustecimento seletivo para algoritmos de PAD. Realizamos campanhas de injeção de falhas para identificar as variáveis de código-fonte mais críticas e apresentamos como selecionar os melhores candidatos para maximizar a relação confiabilidade/sobrecarga.