ALGORITMOS PARA ALOCAÇÃO DE INFRAESTRUTURAS VIRTUAIS CONFIÁVEIS UTILIZANDO GPU

Wilton Jaciel Loch [1], Guilherme Piegas Koslovski[3]

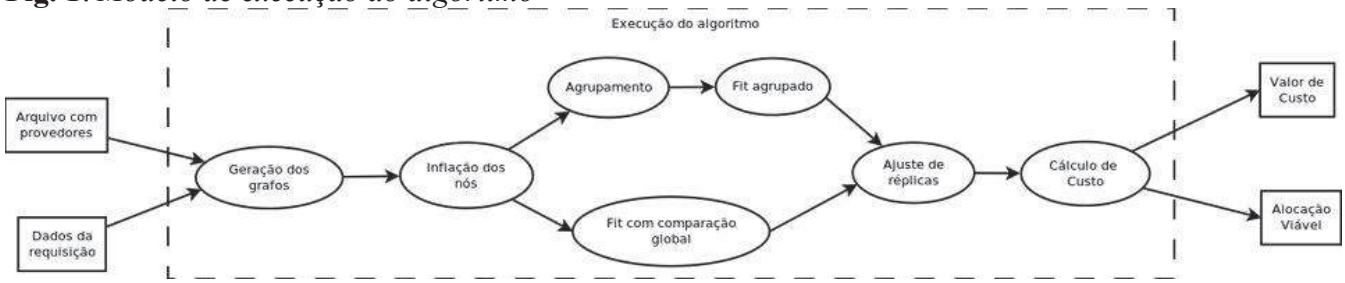
[1] Acadêmico(a) do Curso de Ciência da Computação CCT - bolsista PROBIC/UDESC,
[2] Orientador, Departamento de Ciência da Computação CCT – guilherme.koslovski@udesc.br

Palavras-chave: GPU, infraestruturas virtuais, confiabilidade, IaaS.

Atualmente centenas de serviços populares que são utilizados diariamente dependem de alocação em nuvens computacionais, essas disponibilizam uma infraestrutura com máquinas virtuais para atender o cliente e possibilitar que ele utilize tais recursos para manter o funcionamento de sua aplicação em si. Porém podem haver falhas e problemas técnicos no que diz respeito ao funcionamento de tais nuvens e essas podem ocorrer por diversos fatores afetando de maneira ampla e significativa muitos dos serviços nelas alocados devido à indisponibilidade criada. Os prejuízos causados por falhas podem ser milionários, mesmo que a indisponibilidade dure apenas algumas horas. Aplicações como *Netflix*, *Instagram*, *Pinterest*, *Github*, *Trello*, *Giphy* e *Medium* já foram de alguma forma afetadas por quedas de serviço, prejudicando tanto a empresa quanto os usuários de modo geral. De maneira geral o objetivo do projeto é garantir que as alocações em nuvens sejam confiáveis e que possuam capacidade de sobrevivência por meio de uma modelagem que utilize-se da estrutura de grafos e dos algoritmos conhecidos sobre os mesmos, já adaptados para GPU no *framework* desenvolvido previamente pelo Lucas Nesi [1]. O conceito de uma alocação confiável diz respeito a probabilidade de uma infraestrutura virtual funcionar corretamente em um determinado ponto no tempo. Já o conceito de capacidade de sobrevivência diz respeito a capacidade de uma infraestrutura virtual continuar operacional mesmo após a ocorrência de uma queda no serviço do provedor. Para atingir ambos os conceitos é utilizada uma solução baseada em réplicas para máquinas críticas e no espalhamento da alocação. De modo geral, as réplicas são responsáveis por assumir o lugar de uma máquina crítica caso esta falhe, permitindo que o serviço continue ativo. A quantidade de réplicas necessárias depende das probabilidades de falha de cada zona e de quantas máquinas críticas serão utilizadas. O espalhamento é feito sobre zonas, regiões e provedores distintos, possibilitando assim uma diminuição na probabilidade total de falha do sistema, tendo em vista que ao alocarmos a requisição em diferentes zonas, se houver uma queda de serviço em uma delas não teremos um comprometimento total do sistema. A modelagem encontrada que engloba todos os conceitos acima citados foi utilizar um grafo representando o conjunto de zonas, regiões e provedores sobre o qual os serviços de alocação estão organizados e outro representando uma possível requisição

do cliente, contendo máquinas virtuais e suas características. É então realizada uma alocação do segundo grafo no primeiro, adaptando estratégias de *fit* conhecidas para que cheguemos a uma alocação confiável dentro de um custo aceitável estipulado. No presente momento estão sendo estudadas duas possíveis abordagens para o âmago do processo de alocação, uma baseada na divisão das zonas em grupos de probabilidade e custo e outra baseada em uma comparação global entre as características das mesmas. O processo como um todo possui diversas fases menores que podem ser visualizadas na imagem abaixo:

Fig. 1: *Modelo de execução do algoritmo*



O projeto ainda não foi finalizado, estando ainda em fase de desenvolvimento. Porém, apresenta um desempenho substancialmente bom em relação ao tempo e uso de recursos computacionais. O processo para geração dos dados a respeito da qualidade das alocações de cada abordagem está em desenvolvimento no presente momento.

Referências

[1] Lucas Leandro Nesi, Mauricio Aronne Pillon, Marcos Dias de Assunção, and Guilherme Piegas Koslovski. GPU-accelerated algorithms for allocating virtual infrastructure in cloud data centers. In 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2018), May 2018.