Escalonamento de aplicações com suporte a GPU utilizando contêineres

Resumo Expandido - Disciplina de TCC290009

Helenluciany Cechinel
Estudante do Curso de Engenharia de Telecomunicações

Ederson Torresini
Professor orientador

Semestre 2018-1

Resumo- *A consolidação da computação em nuvem e o desenvolvimento de aplicações baseadas em microsserviços permitem uma melhor escalabilidade de recursos e processos, garantindo a modularidade do sistema. A Coordenadoria de Tecnologia e Comunicação do câmpus São José (CTIC) possui uma cultura de melhoria contínua. Sendo assim, optou-se por uma abordagem de infraestrutura como código, que visa diminuir a redundância de processamento de recursos e a redução dos custos com computadores, utilizando a virtualização por contêineres e escalonamento de aplicações. O intuito do trabalho é aperfeiçoar serviços que já são oferecidos na nuvem privada do Instituto Federal de Santa Catarina campus São José (IFSC-SJ), propondo a otimização de processamento de recursos em CPU através de unidade de processamento gráfico dedicado (GPU).*

Palavras-chave: Contêineres. GPU. Escalonamento. Microsserviços.

1 Introdução

O conceito de computação em nuvem vem se consolidando cada vez mais nas organizações privadas e públicas. É definido como um conjunto de recursos, como capacidade de processamento, armazenamento, conectividade, plataformas, aplicações e serviços disponibilizados pela Internet (TAURION, 2009).

A difusão da computação em nuvem possibilitou a concepção de novas arquiteturas de *software* que visem o provisionamento de recursos computacionais no qual podem executar aplicações distintas. No cenário tecnológico atual, a migração de arquiteturas monolíticas para microsserviços vem aumentando de forma significativa, devido a escalabilidade dos serviços, interoperabilidade entre processos e da infraestrutura sob demanda. De acordo com Carvalho e Anjos (2017), uma arquitetura baseada em microsserviço possui a abordagem de desenvolver uma única aplicação como uma suíte de serviços, cada

um rodando em seu próprio processo e se comunicando através de mecanismos leves, geralmente através de métodos HTTP. Diante do exposto, a literatura de Alves (2017) enfatiza três características fortemente presentes na computação em nuvem: escalabilidade no provisionamento ou liberação de recursos, o modelo de pagamento sob demanda de uso e a virtualização.

Segundo Kang et al. (2017), a virtualização é um dos serviços mais significativos que contribuem para a existência da computação em nuvem. Geralmente dominada pelas tecnologias de máquinas virtuais com uso de *hypervisor*, que utilizam o *software* para emular as funcionalidades de *hardware*, permitindo que vários sistemas operacionais e aplicações diversas sejam executadas em um mesmo ambiente de *hardware físico*. Contudo, a virtualização baseada em contêineres têm surgido para substituir esse conceito de virtualização por *hypervisor*. Um contêiner permite a criação de instâncias de processamentos, sendo possível obter abstração e isolamento de recursos dentro de *namespaces* entre aplicações ou subsistemas. Os ambientes virtuais são criados utilizando recursos e processos presentes no núcleo do próprio sistema operacional. Portanto, a diferença fundamental entre esses dois tipos de virtualização é que o *hypervisor* emula o *hardware* enquanto que o contêiner é a virtualização de aplicações em nível de sistema operacional, agilizando a portabilidade e velocidade das aplicações. Conforme Silva (2017), o uso de contêineres é considerado como uma alternativa leve comparada ao ambiente de *hypervisor*, pois tem como objetivo a redução de sobrecarga do *hardware*, redução de falhas em função da configuração, atualizações, erros de *software* e maior velocidade na inicialização.

O Instituto Federal de Santa Catarina câmpus São José (IFSC-SJ) visa atender a comunidade acadêmica composta por professores, alunos e técnicos administrativos, oferecendo-os acesso as aplicações armazenadas local e remotamente. O IFSC-SJ detém uma estrutura de salas e laboratórios que possuem ferramentas de programação, aplicações de processamento de sinais digitais e simuladores para o uso dos docentes e discentes, dos cursos com ênfase em Telecomunicações. Recentemente, uma nova modalidade de laboratórios experimentais vem sendo estudada e discutida pela instituição de ensino: a proposta de implementação de laboratórios remotos controlados via Internet como apoio ao ensino presencial, podendo também ser usado em ensino a distância (EaD). Os laboratórios remotos permitem que experimentos reais de um laboratório físico sejam controlados remotamente através da nuvem privada. Diante desta realidade e da preocupação no que tange a restrição de acesso aos laboratórios físicos em determinados horários e outros fatores limitantes ao uso dessas aplicações em máquinas locais, que a Coordenadoria de Tecnologia de Informação e Comunicação (CTIC) propôs a adoção da computação na nuvem.

Atualmente, o ambiente de computação é híbrido, ou seja, parte das aplicações rodam em laboratórios físicos e parte na nuvem privada. Isto é, a CTIC disponibiliza os mesmos aplicativos que são executados nas máquinas físicas em ambiente remoto. A justificativa da utilização de nuvem híbrida ainda não teve a devida investigação, mas acredita-se que alguns dos fatores sejam a preocupação do armazenamento dos projetos e experimentos realizados pelos alunos na nuvem, a incompatibilidade de versões dos aplicativos e principalmente a restrição de processamento da aplicações. Visto que muitos

aplicativos podem requerer uso intensivo de CPU, sendo capazes de comprometer o desempenho e eficiência da nuvem devido grande demanda de recursos e necessidade de renderização dessas aplicações.

Para atingir uma boa eficiência global, a CTIC vem trabalhando com o conceito de microsserviços para melhor distribuir os recursos de processamento e memória. A equipe possui a intenção de poder executar todos os serviços ofertados pelo IFSC-SJ através de aplicação como serviço, entregando ao usuário final somente a aplicação desejada via HTTP. Desta forma o objetivo primário do projeto é atender a maior quantidade de acessos simultâneos a essas aplicações, utilizando recursos ociosos da CTIC, como placas de vídeo e a infraestrutura em nuvem do IFSC-SJ. Portanto, este trabalho visa otimizar o escalonamento das aplicações que utilizam poder de processamento intensivo através de recursos dedicados de unidade de processamento gráfico (GPU) ou placas FPGA. Outro objeto de pesquisa deste trabalho é a investigação do o uso de GPU, ao invés de CPU, para aprimorar a experiência de uso dos aplicativos por parte do usuário final, analisando também a melhoria da eficiência do ambiente global de forma a atender uma quantidade maior de usuários na infraestrutura de nuvem.

2 Metodologia

Para o bom desenvolvimento do projeto, se faz necessário a especificação da metodologia do trabalho. A metodologia consiste em 5 etapas fundamentais, as quais estão dispostas nos itens abaixo.

2.1 Levantamento da Infraestrutura Atual

Nesta etapa são analisados os serviços que são fornecidos pela CTIC à comunidade acadêmica. Após análise inicial, surge a necessidade de identificar quais desses serviços requerem intensivo poder de processamento e memória. A partir dos dados obtidos, é feita uma avaliação de quais aplicações poderiam obter suporte ao uso de processamento gráfico dedicado para garantir maior eficiência de processamento global. Estudos de melhoria das aplicações *Web* no quesito eficiência já foram pensados, entretanto, aplicações como as de *Desktop* que demandam uso intensivo de CPU, ainda não foram estudados.

No primeiro momento é realizado um mapeamento da infraestrutura da nuvem privada atual do IFSC, no campus São José. É contabilizado o número de equipamentos ativos e ociosos que são relevantes ao projeto proposto, como por exemplo servidores e placas de vídeo dedicadas. Também é estudado a especificação de cada equipamento, buscando informações de capacidade de processamento, quantidade de núcleos, memória, eficiência energética, largura de banda e sistema operacional.

Para obter melhor compreensão da estrutura de nuvem privada implementada pela CTIC, uma análise da arquitetura e dos serviços ofertados à instituição é realizada por meio de entrevista com membros da equipe e com leitura de documentação dos equipamentos que compõe a estrutura da nuvem do IFSC-SJ, para que assim, seja possível entender melhor o cenário real e levantar os requisitos para o trabalho.

2.2 Levantamento de Requisitos

Esta parte da metodologia é caracterizada pela análise de requisitos do projeto, identificando as métricas que definem o bom desempenho e funcionamento do sistema já existente e elencando as possíveis restrições de implementação da solução proposta. Essa etapa possui a finalidade de definir esses indicadores para melhor atender a necessidade dos usuários da nuvem privada. Dentre essas métricas podem ser destacadas o desempenho de processamento gráfico, desempenho de CPU, memória e outros valores que possam contribuir para definição de uma política de escalonamento dessas aplicações.

Também é proposto um estudo das ementas das disciplinas do curso de Engenharia de Telecomunicações, buscando encontrar as aplicações que rodam na infraestrutura de nuvem mais utilizadas e que possuem maiores números de chamados em aberto, pela decorrência de atraso de execução. Tendo em vista a adição de novos componentes à estrutura do IFSC-SJ, o levantamento dos requisitos é realizado em conjunto com a equipe da Coordenadoria de Tecnologia e Comunicação do IFSC-SJ, afim de remodelar a estrutura da nuvem privada.

2.3 Estudo de Ferramentas para Implantação

Esta etapa do trabalho consiste no aprofundamento do estudo de ferramentas e tecnologias utilizadas para a implementação do escalonamento de aplicações. Portanto, o projeto demanda que sejam feitas leituras conceituais sobre definição e funcionamento de contêineres, escalonamento de processos, aplicações em microsserviços, distribuição de carga, processamento gráfico dedicado e como funciona a concorrência de uso de uma mesma GPU por vários usuários. Além desses estudos conceituais, são realizados laboratórios para melhor compreender o comportamento de tecnologias como *Docker*, *Docker-nvidia*, CUDA e *Kubernetes*.

2.4 Implantação da Solução

O trabalha consiste em aplicar na prática um conjunto de ferramentas e tecnologias para obtenção de maior eficiência no uso de recursos de processamento das aplicações dos usuários finais que utilizam a nuvem privada do IFSC-SJ. Sendo assim, surge a demanda para configuração do *driver* de placa de vídeo da nVidia em que o mesmo deve ser compatível com a imagem do sistema operacional existente. Também é necessário a implantação da plataforma CUDA, cuja a mesma é responsável pelo suporte a computação paralela. Após realização das atividades descritas acima, se faz indispensável a instalação e execução do *Docker Engine* para orquestrar contêineres com suporte a nVidia e CUDA.

Como etapa final, é proposto estabelecer uma política de escalonamento baseada por afinidade de máquina ou grupo de máquina com suporte a aceleração gráfica, utilizando *Kubernetes*. Esse escalonamento permite distribuir a carga entre recursos disponíveis em um *Cluster*, otimizando o processamento na nuvem privada da instituição. A política de escalonamento das aplicações depende da disponibilidade de recurso de processamento e memória. Levando em consideração as métricas levantadas na etapa de levantamento de

requisitos. Métricas de rede são desconsideradas para elaboração dessa política, pois não está sendo tratado de nuvem multi-região.

2.5 Testes e Avaliação

Posterior ao processo de implantação, tem-se a demanda para realização de testes de desempenho e funcionamento da nuvem privada, para garantir que o serviço de escalonamento de aplicações com uso de GPU utilizando contêineres, atenda às necessidades da comunidade acadêmica. As verificações do funcionamento da solução proposta consistem em teste de aplicação, *benchmarking*, testes em nuvem com inserção de erros para verificar a tolerância a falhas do serviço e testes de replicação em outros nós, testando a migração da aplicação para outras máquinas.

Para finalizar a parte de testes, é preciso elaborar códigos que explorem o potencial das aplicações que possibilitem o uso de computação paralela, instigando a utilização de recursos dedicados. Portanto, para auxílio na criação desses testes, se faz necessário a assessoria de um professor com experiência na área de processamento de sinais e renderização. Além dos testes de validação, são elaborados questionários para avaliar a experiência do usuário final em relação ao tempo de resposta das aplicações local e remotas. Desta forma, é possível agrupar dados que justifiquem a necessidade de processamento com uso de GPU para aplicações utilizadas em disciplinas do cursos de Telecomunicações.

3 Considerações Parciais

O intuito da Coordenadoria de Tecnologia e Informação do IFSC-SJ é migrar todos os serviços disponíveis aos usuários para a nuvem privada. A intenção é entregar à comunidade acadêmica todos os serviços oferecidos pela CTIC diretamente por aplicações *Web* ou seja, o usuário só precisaria acessar o navegador e inserir a URL do serviço que está disponível. Diante do exposto, o presente trabalho visa suprir uma demanda futura, e acrescentar o serviço de escalonamento de aplicações com suporte a GPU ao portfólio da equipe. Entretanto, para concluir o objetivo inicial da CTIC, outros trabalhos devem ser desenvolvidos para complementar e justificar este trabalho. Como trabalhos futuros, recomenda-se desenvolver aplicações com aceleração utilizando FPGA e realizar renderização remota das aplicações.

Referências

ALVES, T. H. C. R. *Uma Arquitetura Baseada em Containers para Workflows de Bioinformática em Nuvens Federadas*. [s.n.], 2017. Disponível em: <http://repositorio.unb.br/bitstream/10482/30994/1/2017_TiagoHenriqueCostaRodriguesAlves.pdf>.

CARVALHO, L. S. P.; ANJOS, M. Cézar Lopes dos. *Impacto dos padrões arquiteturais de Micro Serviço e Monolítico no desenvolvimento de softwares*. [s.n.], 2017. Disponível em: <http://repositorio.aee.edu.br/bitstream/aee/50/1/TCC2_2017_02_LucasPedatela_MatheusCesar.pdf>.

KANG, D. et al. *ConVGPU: GPU Management Middleware in Container Based Virtualized Environment*. [s.n.], 2017. v. 2017-September. 301 - 309 p. ISSN 15525244. Disponível em: <http://dx.doi.org/10.1109/CLUSTER.2017.17>.

SILVA, F. H. R. e. *Avaliação de Desempenho de Contêineres Docker para Aplicações do Supremo Tribunal Federal*. [s.n.], 2017. Disponível em: <http://bdm.unb.br/bitstream/10483/17796/1/2017_FlavioHenriqueSilva_tcc.pdf>.

TAURION, C. *Computação em Nuvem. Transformando o mundo da Tecnologia da Informação*. [S.l.: s.n.], 2009. ISBN 9788574524238.