Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Avaliação de Máquinas Preemptáveis nos Provedores de Nuvem Pública Amazon e Google

Jonas P. Soares

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientadora
Prof.a Dr.a Aletéia Patrícia Favacho de Araújo

Brasília
2019

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Avaliação de Máquinas Preemptáveis nos Provedores de Nuvem Pública Amazon e Google

Jonas P. Soares

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof.a Dr.a Aletéia Patrícia Favacho de Araújo (Orientadora)
CIC/UnB

Prof.a Dr.a Maristela Terto de Holanda
CIC/UnB

Prof. MSc. Edward Ribeiro
Senado Federal

Prof. Dr. José Edil Guimarães de Medeiros
Coordenador do Curso de Engenharia da Computação

Brasília, 30 de junho de 2019

Dedicatória

Dedico esta monografia aos meus pais, Olímpia e Edmilson, por darem o suporte necessário para que eu me tornasse a pessoa que sou hoje. Ao meu irmão, Saul, por sempre fazer da nossa convivência amigável desde a infância. Aos meus amigos, pelos momentos de descontração que aliviaram o peso da rotina universitária. Por fim, dedico à Karolyne Antunes, pelo apoio incondicional em todas as esferas da vida e por permitir que construamos juntos um projeto de vida.

Agradecimentos

Agradeço à Profª. Drª. Aleteia, pela dedicação aos seus alunos e pelas orientações prestadas ao longo de mais de um ano de acompanhamento. Agradeço à Universidade de Brasília; a universisade pública que é mais que um lugar de capacitação profissional, mas um lugar formação cidadã e instrumento de transformação social, para o qual desejo voltar em breve.

Além disso, agradeço às instituições de fomento à pesquisa, em especial ao CNPq e à FAP-DF, por acreditarem nos pesquisadores presentes nos cursos de graduação; foi o apoio dessas instituições que possibilitou a descoberta do meu interesse pela pesquisa e docência.

Por fim, agradeço à educação pública e à instituição democrática brasileira, as quais se encontram fragilizadas no momento político em que este trabalho é desenvolvido, e precisam ser protegidas com afinco.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

No contexto contemporâneo, no qual diversas empresas e organizações possuem grandes demandas por recursos computacionais, cada vez mais provedores de nuvem pública surgem no mercado. Diante disso, a escolha do serviço e do provedor mais adequados se torna um desafio não trivial para usuários. Nesse contexto, este trabalho propõe uma análise comparativa sobre máquinas preemptáveis oferecidas por provedores de nuvem pública, as quais podem ser finalizadas em situações onde seus recursos computacionais são necessários em outras tarefas do provedor do serviço. Para isso, são executados testes experimentais em instâncias oferecidas pelos provedores, utilizando *benchmarks*. O trabalho conclui, a partir dos resultados de custo e de performance obtidos, quais instâncias, provedores e regiões são mais indicados para cargas de trabalho similares aos *benchmarks* executados.

Palavras-chave: Nuvem Computacional, *Benchmarks*, Custo-benefício, Máquinas Preemptáveis, Instâncias Spot, Amazon, Google

Abstract

In the contemporary context, in which several companies and organizations have great demands for computational resources, more and more public cloud providers arise on the market. Therefore, choosing the right service and provider becomes a non-trivial challenge for users. In this context, this paper proposes a comparative analysis on preemptible machines offered as a service by public cloud providers. For that goal, experimental tests are executed in instances offered by the providers, using benchmarks. The work concludes, from the results of cost and performance obtained, which instances, providers, and regions are best suited for benchmarked workloads.

Keywords: Cloud Computing, Benchmarks, Cost–benefit, Preemptible VMs, Spot Instances, Amazon, Google

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AWS *Amazon Web Services.*

EBS *Elastic Block Store.*

EC2 *Elastic Compute Cloud.*

GCE *Google Compute Engine.*

GCP *Google Cloud Platform.*

GPU *Graphics Processing Unit.*

HDD *Hard Disk Drive* ou Unidade de Disco Rígido.

I/O *Input e Output* ou Entrada e Saída.

IaaS *Infrastructure as a Service.*

JRE *Java Runtime Environment.*

JVM *Java Virtual Machine.*

NIST *National Institute of Standards and Technology.*

PaaS *Platform as a Service.*

QoS *Quality of Service.*

SaaS *Software as a Service.*

SLA *Service Level Agreement.*

SSD *Solid-State Drive* ou Unidade de Estado Sólido.

TIC Tecnologia da Informação e Comunicação.

VM *Virtual Machine* ou Máquina Virtual.

Capítulo 1

Introdução

A sociedade contemporânea é caracterizada pela presença difusa da Tecnologia da Informação e Comunicação (TIC), em seus diversos aspectos. Com a popularização da Internet, a comunicação entre seres humanos e sistemas interativos tomou grandes proporções, gerando um grande volume de dados, e com isso a necessidade de processamento em larga escala. Nesse cenário, surgem plataformas nas quais os usuários têm acesso ao poder computacional desejado com um investimento reduzido: as nuvens computacionais. Com a nuvem, organizações oferecem infraestrutura de recursos computacionais como serviço, movimentando um mercado cuja receita projetada para 2019 é de mais de 200 bilhões de dólares, segundo a companhia *Gartner, Inc.* [1].

Em um mercado amplo como esse, onde diversos provedores ofertam serviços similares, identificar as vantagens oferecidas por cada provedor é de grande valia para potencias usuários. Visando ampliar o mercado consumidor e maximizar lucros, diversos provedores oferecem modelos de serviço baseados em máquinas preemptáveis, também chamadas de instâncias preemptáveis, nas quais o valor cobrado é reduzido drasticamente em detrimento da confiabilidade oferecida aos usuários. Este modelo pode ser de grande interesse para determinados perfis de clientes, os quais são habilitandos a executar tarefas com custo reduzido. A comparação entre diferentes provedores, entretanto, configura uma tarefa não trivial para potenciais clientes desses serviços, uma vez que cada empresa implementa a oferta de máquinas preemptáveis de maneira singular.

Nesse contexto, este trabalho propõe uma análise comparativa de máquinas preemptáveis oferecidas por provedores de nuvem pública, especificamente sobre os provedores Amazon e Google. Para isso, são executados testes experimentais em diversas instâncias oferecidas pelos provedores - testes estes baseados em aplicações desenvolvidas especialmente para análises de performance, chamadas *benchmarks*. Com isso, o trabalho busca investigar, sob uma perspectiva de custo-benefício, eventuais vantagens comparativas observadas nos resultados produzidos por cada instância avaliada.

1.1 Objetivos

O objetivo geral deste trabalho é investigar os serviços de máquinas preemptáveis oferecidos por provedores de nuvem pública e avaliar custo e performance desses serviços, partindo do interesse do usuário de usufruir de um serviço adequado pelo menor preço. Para atingir o objetivo geral, foram definidos os seguintes objetivos específicos:

- Realizar estudo sobre os modelos de oferta e precificação aplicados a serviços de máquinas preemptáveis em nuvem pública;
- Aplicar *benchmarks* em máquinas preemptáveis oferecidas por provedores de nuvem variados, produzindo resultados quantitativos;
- Desenvolver e interpretar gráficos que ilustrem o custo-benefício mensurado nos resultados;
- Produzir conclusões que indiquem quais tipos de instâncias apresentam melhor custo-benefício dentre as avaliadas.

1.2 Metodologia

Para o desenvolvimento deste trabalho, a metodologia utilizada se baseou na divisão de tarefas em etapas. a primeira etapa consistiu em uma pesquisa bibliográfica sobre computação em nuvem e máquinas preemptáveis, para a fundamentação teórica e contextualização do leitor.

A etapa seguinte foi a definição dos provedores cujas máquinas preemptáveis seriam contempladas pelo experimento deste trabalho. Essa decisão foi tomada com base no quadrante mágico divulgado pela companhia Gartner e referente ao ano de 2018 (Figura 1.1). Segundo análise da companhia [2], o provedor Amazon manteve sua liderança no mercado de IaaS, o que motivou sua inclusão no escopo deste trabalho. Como segundo provedor, o Google foi selecionado devido à sua recente inclusão no quadrante de líderes *(leaders)*. Ainda que apresente grande relevância segundo a análise pela Gartner, a Microsoft não foi contemplada por esta análise devido ao seu ingresso ainda muito recente no mercado de máquinas preemptáveis, datado de 2017 [3] - Amazon e Google trouxeram suas versões do serviço já em 2009 e 2015, respectivamente. Em suma, este trabalho propõe uma análise sobre a principal empresa do mercado de nuvem e sobre uma concorrente de crescimento expressivo, com mais de 3 anos de experiência na oferta de máquinas preemptáveis.

A terceira etapa do trabalho consistiu na definição de métodos para a comparação entre as máquinas preemptáveis dos provedores. Para tal, foram utilizadas aplicações desenvolvidas para avaliação de performance, chamadas *benchmarks*. Assim sendo, buscou-se

Figura 1.1: Quadrante Mágico para serviços do tipo IaaS em 2018.

avaliar a execução de cargas de trabalho com diferentes comportamentos. Para isso, foram executados testes com os *benchmarks* SPECjvm2008 e Sysbench *fileio* - caracterizados por um processamento intensivo e por um grande volume de operações de leitura e escrita, respectivamente.

Em seguida, foram preparados ambientes de teste em ambos os provedores, de maneira que as diferenças entre esses ambientes fossem mínimas. Com isso, os resultados obtidos na execução dos testes poderiam, conforme desejado, refletir os impactos causados pela infraestrutura dos provedores. Preparados os ambientes, diversas máquinas foram instanciadas em ambos os provedores, produzindo uma ampla gama de resultados a partir da execução dos *benchmarks*.

A quinta e última etapa do trabalho consistiu na análise dos resultados, possibilitando assim a interpretação e produção de conclusões relevantes.

1.3 Estrutura do Trabalho

Este trabalho se encontra estruturado em 4 capítulos, além deste. O Capítulo 2 apresenta uma definição de nuvem computacional e uma análise dos principais modelos de precificação utilizados em nuvens públicas, apresentando os conceitos de instâncias dedicadas e preemptáveis. O Capítulo 3 descreve a proposta de análise deste trabalho, explorando características dos provedores avaliados e descrevendo os *benchmarks* com os quais as máquinas foram testadas. O Capítulo 4 explora graficamente os resultados obtidos, apontando recomendações de máquinas preemptáveis sob a perspectiva de cada *benchmark* utilizado. Por fim, o Capítulo 5 descreve as conclusões obtidas neste trabalho e apresenta sugestões para trabalhos futuros acerca da proposta.

Capítulo 2

Computação em Nuvem

A seção 2.1 deste capítulo apresenta uma definição de nuvem computacional, seus variados modelos e suas principais características no paradigma de computação distribuída. A seção 2.2 aborda o desafio da precificação e sua relevância para o sucesso de um mercado de serviços. Por fim, a seção 2.3 descreve os principais modelos comerciais utilizados pelos provedores de nuvem para lidar com esse desafio.

2.1 Definição e Modelos de Computação em Nuvem

Giordanelli [4] escreveu, em 2010, que a ideia precursora de nuvens computacionais surgiu a partir de uma evolução do conceito de *grids* computacionais. Um *grid* consiste em um sistema que coordena recursos de maneira descentralizada, utilizando protocolos e interfaces padronizadas para entregar poder computacional com uma qualidade não trivial. Enquanto os *grids* possuem foco em uma infraestrutura que entrega recursos de armazenamento e computação, as nuvens visariam oferecer recursos e serviços de maneira mais abstrata, partindo de um olhar orientado para o mercado.

Sob a definição de Buyya [5] em 2009, as nuvens são caracterizadas pelo forte suporte a virtualização e o provisionamento de recursos sob demanda. Virtualização é a configuração de um hardware de maneira que seus recursos possam ser encapsulados em diversas máquinas virtuais (instâncias) isoladas. Dessa maneira é possível, por exemplo, decompor *mainframes* e *datacenters* em instâncias customizadas de pequeno e médio porte. Buyya [5] acreditava que a computação se tornaria um serviço de utilidade diária como telefonia e energia elétrica, tendo a nuvem computacional um papel fundamental na popularização dos serviços computacionais.

Em 2011 o *National Institute of Standards and Technology (NIST)* [6] publicou uma recomendação que se tornou a principal definição da literatura, conceituando computação

em nuvem como um modelo de computação distribuída baseada em cinco características essenciais:

- Autosserviço sob demanda: O consumidor pode provisionar recursos de computação, como tempo de servidor e armazenamento em rede, por conta própria e sem necessitar intervenção humana dos provedores de serviços;
- Amplo acesso por rede: Os recursos computacionais estão disponíveis através da rede e são acessados através de mecanismos padronizados, os quais promovem o uso por dispositivos clientes de diversas plataformas (como smartphones, tablets, laptops ou desktops);
- Recursos virtualizados: Os recursos de computação de cada provedor são concebidos para servir vários clientes, cada um com diferentes recursos virtuais alocados dinamicamente. Armazenamento, processamento, memória, largura de banda de rede e máquinas virtuais são alguns exemplos de recursos disponibilizados dessa maneira;
- Elasticidade rápida: um importante conceito que surgiu no paradigma de nuvem, elasticidade diz respeito ao provisionamento ou liberação de recursos em tempo real, de acordo com a necessidade do usuário [7]. Ela pode ser implementada tanto pelo redimensionamento das máquinas virtuais alocadas ao usuário - chamada elasticidade vertical; quanto pela adição de novas máquinas virtuais ao serviço prestado - chamada elasticidade horizontal [7]. Essa propriedade possibilita ao consumidor a aparência de que os recursos são ilimitados e podem ser provisionados a qualquer momento e em qualquer quantidade;
- Serviço mensurável: Os sistemas na nuvem controlam o uso dos recursos através de medições em um nível de abstração apropriado para o tipo de serviço (como armazenamento, processamento, comunicação de rede e contas de usuário ativas). A utilização de recursos pode ser monitorada, controlada e informada, gerando transparência tanto para o fornecedor como para o consumidor do serviço utilizado.

Em 2014, Hashem *et al.* [8] apresentou a nuvem como um modelo que permitisse acesso conveniente e sob demanda a recursos computacionais que pudessem ser provisionados rapidamente e disponibilizados com esforço de gerenciamento mínimo. Ainda que tenham características únicas, todas as definições acima citadas trazem ideias relacionadas à disponibilidade de recursos virtualizados, monitorados e acessíveis via rede.

Diante das características acima citadas, é possível elencar uma série de vantagens no uso da tecnologia de nuvens computacionais. Algumas delas são [9]:

- Sensação de recursos computacionais infinitos disponíveis sob demanda, eliminando a necessidade de usuários estimarem antecipadamente a escala de recursos necessários;
- Ausência de barreira de entrada para usuários de nuvens, possibilitando que pequenas empresas e organizações possam iniciar projetos de baixo custo inicial e aumentem seu consumo de recursos na medida que julgarem necessário;
- Possibilidade de pagar por recursos computacionais por seu uso efetivo, com cobranças em unidades de tempo como dias, horas e segundos. Com isso, uma boa gestão de recursos pode reduzir desperdícios e cortar custos por parte do usuário.

Muitas das publicações com definições de nuvem computacional trazem conceitos relacionados a possíveis categorias das nuvens. Existem três principais modelos de serviço de nuvem segundo o NIST [6], sendo que em cada um deles o usuário possui um nível de abstração diferente sobre a máquina virtual:

- ***Infrastructure as a Service* (IaaS)**: o provedor disponibiliza recursos computacionais (processamento, armazenamento, entre outros) de maneira direta. O usuário é capaz de gerenciar e controlar sua máquina virtual de maneira independente, incluindo sistemas operacionais e aplicações executadas na mesma;
- ***Platform as a Service* (PaaS)**: o cliente utiliza uma plataforma que pode ser utilizada por meio de bibliotecas, linguagens de programação, entre outros. O usuário é responsável pela gestão de aplicações executadas na máquina virtual. Elementos de infraestrutura como rede, sistemas operacionais e discos de armazenamento são definidos e gerenciados pelo provedor;
- ***Software as a Service* (SaaS)**: o usuário faz o uso direto das aplicações oferecidas pelo provedor, as quais são acessadas por meio de uma interface (*web browser*, aplicativos). O nível de controle do usuário está limitado a configurações específicas de usabilidade, sem acesso direto às máquinas virtuais utilizadas pelo serviço.

Em relação aos tipos de nuvem, segundo NIST [6], existem também diferentes modelos de implementação. Esta classificação leva em conta os objetivos da organização que mantém a nuvem e o perfil de seus usuários:

- **Nuvem pública**: sua infraestrutura é provisionada para uso aberto de um público-alvo genérico. Pode ser gerenciada/operada por uma ou mais organizações com ou sem fins lucrativos;

- **Nuvem privada**: provisionada para uso exclusivo de uma única organização, em caráter interno. Essa solução é recomendada, principalmente, em cenários onde a privacidade de dados é um problema crítico para o usuário;
- **Nuvem comunitária**: provisionada para uso de um grupo de pessoas/organizações com valores em comum. O seu controle e a operação podem ser feitos por organizações pertencentes à comunidade, agentes externos ou um misto entre eles;
- **Nuvem híbrida**: composta de duas ou mais nuvens com diferentes modelos de implementação, de maneira que tais nuvens se comuniquem por meio de protocolos e/ou portabilidade por software.

Assim, ao identificar o surgimento de um novo mercado com as nuvens, organizações que já possuíam experiência em gerenciamento de *datacenters* identificaram oportunidades de negócio e se lançaram no mercado como provedoras de nuvens públicas, oferecendo VMs e serviços de armazenamento a partir de uma cobrança monetária. Este é atualmente o modelo de nuvem mais difundido no mercado e mais popular entre empresas [10].

Dessa forma, diferentes provedores de nuvem pública têm apresentado diversas formas de mensurar a cobrança pelo uso de seus recursos. Visando um melhor entendimento desta diversidade, a seção 2.2 aborda esse assunto em detalhes.

2.2 Precificação de Nuvens

Segundo Al-Roomi *et al.* [11], o principal objetivo de um provedor de nuvem computacional típico é maximizar seus rendimentos, enquanto o principal objetivo de seus clientes é utilizar um serviço de melhor qualidade possível por um preço razoável. Diante desse conflito de interesses, é o processo de precificação que determina aquilo que um provedor receberá do usuário final em compensação por seus serviços prestados, caracterizando o custo final para um usuário. Weindhardt *et al.* [12] alegaram que o sucesso da nuvem pública no mercado somente poderia ser atingido através de técnicas de precificação adequadas por parte dos provedores. Este valor final pode ser estabelecido de algumas maneiras no mercado de serviços:

- **Fixa**: precificação realizada a partir de um valor pré-estabelecido em um catálogo de valores, no qual o pagamento deste valor garante o uso do recurso por um período ilimitado (*pay once*), ou por um intervalo de tempo (*pay-per-use*);
- **Dinâmica**: precificação cujo valor final é reajustado dinamicamente devido a características do serviço, quantidade de volumes adquiridos ou preferências do cliente;

- **Dependente do mercado**: precificação ajustada em tempo real devido a condições de mercado como leilões, nível de demanda do serviço e recursos disponíveis.

Custos iniciais, condições dos recursos oferecidos, custos de manutenção do provedor e personalização do serviço são alguns dos elementos relevantes para a determinação do preço final de serviços de nuvem. É importante ainda que fatores referentes a provedores concorrentes sejam ponderados, buscando oferecer melhores condições, custos ou outras vantagens comparativas para seus usuários [11].

Indo além da análise monetária, os *Service Level Agreement* (SLA) são valiosos artefatos para o exercício de comparação entre diferentes serviços e provedores. Os SLAs são termos de acordo que especificam os níveis de qualidade (QoS) oferecidos por um provedor, e as garantias mínimas entregues aos usuário de seu serviço, constituindo assim um importante material de referência para eventuais auditorias sobre os serviços. Tipicamente, SLAs de nuvens especificam as métricas que deverão atingir durante a prestação do serviço contemplado. Alguns exemplos de garantias são tempo mínimo de disponibilidade (em porcentagem), tempo de resposta limite (em milissegundos) e recuperação a possíveis desastres [13].

Sob a perspectiva de usuários de Tecnologia da Informação e Comunicação (TIC), ter acesso aos recursos computacionais necessários para suas operações com custo otimizado se tornou uma prioridade que resulta em cada vez mais organizações aderindo ao mercado de nuvens públicas. Além disso, o reconhecimento de demandas variadas dentro das organizações tem levado grande parte dessas organizações a diversificar também os serviços de nuvem contratados, com muitas delas utilizando nuvens de mais de um provedor [10].

Diante desse cenário, a sessão 2.3 deste capítulo apresenta os principais modelos de precificação utilizados pelos provedores de nuvem, cujo conhecimento é de importante valor para que usuários façam melhores escolhas quanto aos serviços de nuvem oferecidos no mercado.

2.3 Instâncias Dedicadas e preemptáveis

Diante da complexidade do cenário apresentado anteriormente, a maioria dos provedores de *IaaS* adotou um modelo de precificação fixo, no qual diferentes configurações de máquinas virtuais (VMs) são oferecidas aos consumidores, e cada uma delas possui um preço pré-determinado por unidade de tempo de uso. Essas instâncias são de uso dedicado *(on-demand)* e contam com SLAs que especificam o coeficiente mínimo de disponibilidade entregue ao cliente do serviço, sendo esse comumente acima de 99,9% [14] [15].

Este modelo atribui ao provedor o desafio de definir um preço fixo no ponto de equilíbrio em um mercado onde a demanda é altamente volátil, ou seja, em períodos de su-

butilização da capacidade máxima da nuvem, preços mais baixos poderiam trazer maior mercado consumidor e otimizar rendimentos; em contrapartida, oportunidades lucrativas são perdidas em momentos de demanda superior à capacidade de sua nuvem. Em ambas as situações, o ônus da precificação imprecisa fica com o provedor do serviço, uma vez que o preço foi pré-fixado anteriormente [16].

Diante do exposto, identificando a possibilidade de diminuir perdas causadas pelas desvantagens do modelo fixo e diversificar seus catálogo de serviços, alguns provedores iniciaram projetos de precificação dependente do mercado - lançando então as instâncias preemptáveis [17] [18]. Esse tipo de serviço oferece ao consumidor a execução de sua máquina virtual por um preço reduzido em detrimento da confiabilidade do serviço oferecido. A disponibilidade de instâncias nesse modelo não conta com as mesmas garantias do modelo fixo, uma vez que seus repositórios são criados a partir dos recursos não utilizados por instâncias dedicadas [16]. Tipicamente, a execução de máquinas preemptáveis custa apenas uma fração do valor cobrado por máquinas dedicadas com a mesma configuração. Desta maneira, o modelo se torna interesssante para o mercado consumidor e pode gerar receita ao provedor a partir de recursos previamente ociosos [19]. A Figura 2.1 mostra, a título de exemplo, um gráfico evolutivo de preços praticados pelo provedor Amazon para instâncias do tipo *a1.large* entre 19 de março de 2019 e 17 de junho do mesmo mês, no qual é apresentada uma variação considerável entre o custo de máquinas dedicadas e preemptáveis.

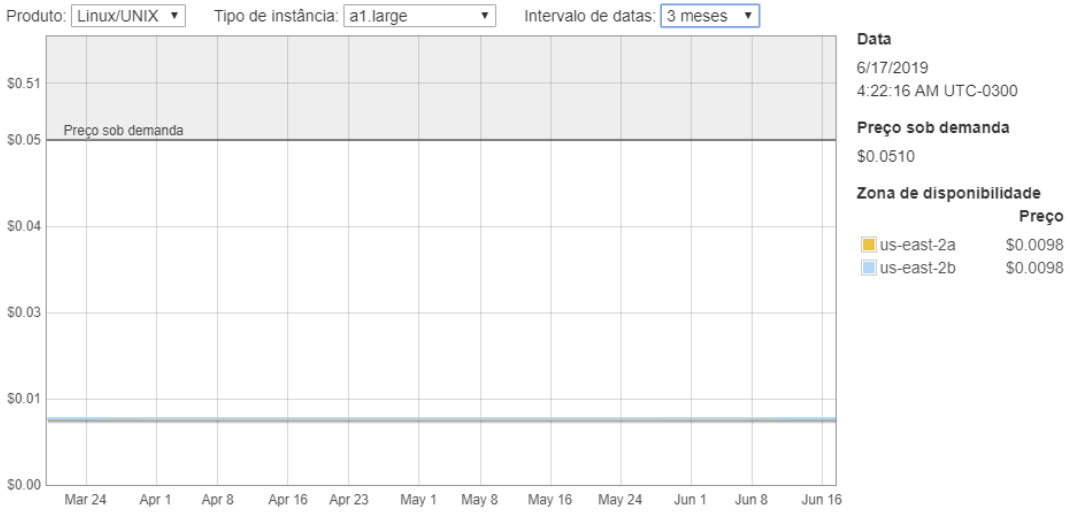

Figura 2.1: Histórico de preços para instâncias Amazon *a1.large*[1].

Dadas as características apresentadas anteriormente, instâncias preemptáveis são recomendadas para demandas computacionais tolerantes a interrupções inesperadas, e que

[1]Captura realizada em 17 de junho de 2019

não sejam sensíveis a queda de performance. Tais instâncias também são uma boa alternativa para usuários que buscam reduzir despesas sem perder acesso a recursos extras quando necessário. Esse modelo de instância já é amplamente difundido para cargas de processamento de alta performance e em vários *workflows* científicos como modelagem climática, *design* farmacêutico e análise de dados [20] [21].

É relevante pontuar que este é um domínio de serviço recente, no qual mesmo as nuvens públicas mais consolidadas estão em atividade há pouco mais de 10 anos [4]. Assim como fizeram as instâncias preemptáveis, novas implementações de modelos de precificação podem se estabelecer no mercado futuramente.

2.4 Considerações finais

Neste capítulo foram apresentadas as principais características da computação em nuvem e de seus modelos. Além disso foi descrito o processo de precificação de serviços e como ele é aplicado pelos provedores de nuvem pública. Por fim, foi apresentado o modelo preemptivo de instâncias, suas motivações as e vantagens em relação ao modelo dedicado. O capítulo seguinte propõe uma análise do custo-benefício entregue por instâncias preemptáveis de diferentes provedores, descrevendo os provedores avaliados, as aplicações utilizadas na análise e trabalhos relacionados da literatura.

Capítulo 3

Ambiente Experimental

Diante dos conceitos apresentados no capítulo anterior, este capítulo propõe uma análise comparativa sobre serviços *IaaS* de diferentes provedores, explorando especificamente a oferta de máquinas preemptáveis desses provedores. A Seção 3.1 apresenta a proposta deste trabalho e a metodologia adotada para aquisição de dados utilizados na análise. A Seção 3.2 descreve os provedores analisados neste trabalho, explorando algumas características particulares no que se refere à oferta de máquinas preemptáveis. A Seção 3.3 apresenta os programas utilizados para a produção de resultados mensuráveis do desempenho das máquinas testadas neste trabalho, os quais são denominados *benchmarks*. Por fim, a Seção 3.4 apresenta obras da literatura que permeam conceitos interessantes a este trabalho.

3.1 Análise Proposta

Diante das vantagens oferecidas aos usuários de nuvem computacional, este trabalho propõe uma análise comparativa e experimental sobre os serviços de infraestrutura computacional (IaaS) de nuvem pública, especificamente sob a perspectiva de máquinas virtuais preemptáveis. O objetivo desta análise é produzir resultados e conclusões que facilitem ao usuário identificar o provedor e o tipo de máquina mais adequados para as suas necessidades.

Para Li *et al.* [22] é importante que a comparação seja por meio de métricas que desconsiderem detalhes de implementação de cada provedor, e tenham foco no desempenho de ponta a ponta dos serviços. Os autores propõem uma série de métricas adequadas para descrever a performance de nuvens computacionais - muitas das quais são baseadas em experimentos com aplicações adequadas para avaliação de desempenho de sistemas, também chamadas de *benchmarks*.

O trabalho aqui proposto explora, especificamente sob a perspectiva de máquinas preemptáveis, quatro das métricas propostas por Li *et al.* [22]: tempo de execução do *benchmark*, custo por *benchmark*, taxa de transferência para dispositivos de armazenamento, e custo com operações de leitura e de escrita nesses dispositivos [22]. Essas características são avaliadas com base em resultados obtidos nos *benchmarks* SPECjvm2008 [23] e Sysbench *fileio* [24] em máquinas preemptáveis dos provedores Amazon e Google, que serão melhor apresentados nas subseções a seguir.

3.2 Provedores Avaliados

Em um mercado em plena expansão como o de nuvem comptacional, no qual o crescimento é de mais de 50% ao ano [25], diversas empresas estão frequentemente inovando em tipos de serviço e estratégias de precificação. Diante disso, este trabalho explora o serviço de máquinas preemptáveis de dois dos quatro principais fornecedores de nuvem do tipo de serviço IaaS, sendo estes Amazon e Google. A primeira se destaca com mais de 30% do mercado mundial, e o segundo apresenta uma forte concorrência, ainda que tenha se lançado como provedor tardiamente [26]. As subseções a seguir exploram brevemente a plataforma de cada provedor, precificações aplicadas sobre suas instâncias preemptáveis, e as estratégias adotadas para lidarem com as interrupções.

3.2.1 Provedor Amazon

Pioneira no mercado de nuvem pública, a companhia Amazon [27] iniciou a oferta IaaS em 2006 através de uma versão *beta* do *Elastic Compute Cloud* (EC2), o qual contava com apenas a região *us-east-1*, localizada na Virgínia do Norte-USA [28]. Atualmente, o provedor conta com 21 regiões geográficas (Figura 3.1).

Ao longo dos anos, o provedor diversificou cada vez mais seus serviços de nuvem, compondo assim a plataforma *Amazon Web Services* (AWS) (Figura 3.2). Além do EC2, este trabalho também leva em consideração o serviço *Elastic Block Store* (EBS), responsável por oferecer volumes de armazenamento em blocos para instâncias EC2 [29].

Explorando a oferta de máquinas virtuais, o provedor Amazon disponibiliza uma ampla lista de configurações de hardware para atender a diferentes casos de uso. Os tipos de máquinas virtuais, também chamados de famílias de instâncias, consistem em determinadas combinações de CPU, memória, armazenamento e capacidade de rede, disponibilizados em diferentes tamanhos. As famílias de instâncias são divididas em máquinas de uso geral, com configurações genéricas, e máquinas especializadas, cuja configuração é otimizada para um recurso específico (processamento, memória ou armazenamento) [30].

Figura 3.1: Mapa da infraestrutura global da AWS.

Figura 3.2: Logo da plataforma *Amazon Web Services*.

A Figura 3.3 apresenta as configurações de instâncias da família M5 como exemplo, considerada uma família de uso geral, cujos dispositivos de armazenamento são gerenciados exclusivamente através do serviço *Elastic Block Store* (EBS), oferecido na plataforma AWS.

Quanto aos modelos de precificação, a AWS foi a primeira plataforma de nuvem a oferecer máquinas virtuais no modelo preemptável, chamadas aqui de Instâncias Spot *(Spot Instances)*, e amplamente estudadas pela literatura [16, 19, 31, 32]. Grande parte das instâncias acima citadas são oferecidas tanto no modelo de precificação de uso dedicado quanto no modelo de instâncias preemptáveis, de maneira que apenas algumas famílias são contempladas pela análise deste trabalho.

A classificação Spot caracteriza um grupo de máquinas cuja disponibilidade é definida por um mercado de leilões, de maneira que os preços das instâncias (chamados preços Spot) são ajustados dinamicamente com base na oferta e na demanda de cada tipo de

Modelo	vCPU*	Mem (GiB)	Armazenamento (GB)	Largura de banda dedicada do EBS (Mbps)	Performance de rede (Gbps)
m5.large	2	8	Somente EBS	Até 3.500	Até 10
m5.xlarge	4	16	Somente EBS	Até 3.500	Até 10
m5.2xlarge	8	32	Somente EBS	Até 3.500	Até 10
m5.4xlarge	16	64	Somente EBS	3.500	Até 10
m5.12xlarge	48	192	Somente EBS	7.000	10
m5.24xlarge	96	384	Somente EBS	14.000	25
m5.metal	96*	384	Somente EBS	14.000	25

Figura 3.3: Exemplos de instâncias do tipo M5.

instância. O principal benefício oferecido pela precificação Spot é a economia em relação ao modelo fixo, uma vez que usuários que ofereçam um valor igual ou maior que o preço Spot vigente ganham acesso às máquinas virtuais em questão [32]. Em suma, o modelo Spot faz com que o preço praticado pelo provedor se aproxime do ponto de equilíbrio de mercado mais consistentemente que no modelo dedicado, de precificação fixa.

3.2.2 Provedor Google

O primeiro serviço de nuvem oferecido pelo Google como provedor de nuvem foi em 2008 e se deu pela oferta de um serviço do tipo PaaS: o *Google App Engine* [33] é uma plataforma de desenvolvimento e hospedagem de aplicações web em *datacenters* gerenciados pelo provedor, oferecendo alocação de recursos adicionais automatizada em situações de demanda adicional [28]. O ingresso da empresa no mercado de IaaS só ocorreu em 2012, com o lançamento do *Google Compute Engine* (GCE) [34]. Esse novo serviço disponibiliza o provisionamento de máquinas virtuais sob demanda na infrastrutura global do Google, onde são executados diversos serviços da companhia como Gmail, YouTube, buscador, entre outros [35]. Atualmente, a plataforma de nuvem Google é denominada *Google Cloud Platform* (GCP) (Figura 3.4) e conta com ampla diversidade de serviços ofertados em 20 regiões distribuídas globalmente (Figura 3.5).

Diferentemente do praticado pela AWS com o EC2 e EBS, tanto a gerência de instâncias quanto de discos é feita, no provedor Google, por meio do GCE. Com isso, o monitoramento de desempenho e custos de ambos os tipos de recurso ficam centralizados, simplificando a interação entre a plataforma e o usuário.

Figura 3.4: Logo da *Google Cloud Platform*.

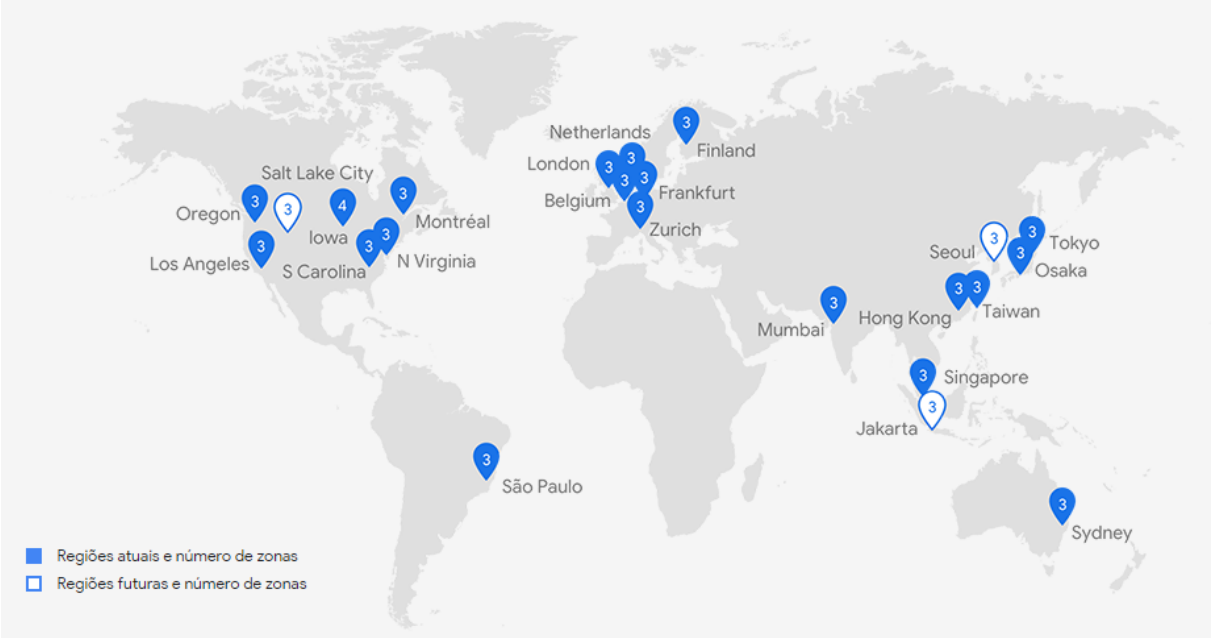

Figura 3.5: Mapa da infraestrutura global da Google.

Diferente do modelo AWS, que oferece diversas famílias de instâncias com configurações de hardware únicas e detalhadas para cada uma delas, a Google apresenta tipos de instâncias cuja variação está apenas com base no volume de memória RAM alocada, geração e quantidade de vCPUs Intel. Além das máquinas padrões, a Google disponibiliza a possibilidade de definir a quantidade de recursos alocados para uma instância de maneira simples e intuitiva, conforme mostra a Figura 3.6. Para uma execução de custo reduzido, o provedor oferece ainda máquinas virtuais de núcleo compatillhado, onde o recurso físico é compartilhado entre diversas instâncias [36].

O preço final de uma instância de uso dedicado do GCE é calculado pela plataforma dinamicamente, de acordo com as definições escolhidas pelo usuário na configuração da máquina. É utilizado um modelo de precificação fixa, no qual o valor calculado é apresentado para o usuário em tempo real, de maneira que ele conheça suas despesas antes mesmo da adesão ao serviço.

As suas instâncias preemptáveis, chamadas pelo provedos de máquinas preemptivas, operam em um modelo de precificação muito similar às instâncias de uso dedicado, ou

Figura 3.6: Provisão de recursos no *Google Cloud Platform*.

seja, não operam com preços de leilão. No modelo praticado pelo Google, essas instâncias possuem preço por hora de execução consideravelmente mais baixo, mas contam com algumas limitações. Primeiramente, elas não são contempladas por nenhuma SLA do provedor, de maneira que não há compensação por períodos de indisponibilidade das instâncias.

No que se refere a interrupções, instâncias preemptivas sempre serão preemptadas após 24 horas contínuas de execução. Quando preemptada, a instância executa a rotina de finalização definida pelo usuário (se houver) e seu estado é alterado para *terminated*, no qual não há uso dos recursos computacionais ou cobrança por horas de instância [18]. Para retomar o estado *running*, onde os recursos da instância estão disponíveis, é nescessário que o usuário dê um comando *start* na plataforma. Além dessa interrupção após 24 horas, máquinas dessa categoria podem estar indisponíveis ou ser preemptadas após intervalos de tempo menores, devido ao esgotamento de recursos do GCE ou outros eventos do sistema.

3.3 Benchmarks utilizados

No contexto deste trabalho, *benchmark* [37] caracteriza a execução de programas a fim de avaliar a performance relativa de um hardware, software ou combinação de ambos. Tipicamente, esse processo envolve a produção de resultados experimentais, cujo significado está intimamente vinculado à comparação entre diferentes cenários.

Além de nomear o processo, a expressão *benchmark* também pode ser utilizada para referenciar programas desenvolvidos especialmente para avaliações desse tipo. Neste trabalho, máquinas preemptáveis dos provedores Amazon e Google são avaliadas sob perspectiva de dois *benchmarks* de código aberto, os quais são o SPECjvm2008 e o Sysbench *fileio*, apresentados nas próximas seções.

Tabela 3.1: Benchmarks da coleção SPECjvm2008.

Benchmark	**Descrição**
Compiler	Compila um conjunto de arquivos .java.
Compress	Realiza uma compressão de dados utilizando um método similar ao LZW. Em suma, a carga de trabalho encontra *strings* comuns e as substitui pelo código de uma variável.
Crypto	Executa cifragem e decifragem utilizando protolocos AES, DES e RSA, utilizando entradas de tamanho variável.
Derby	Executa uma carga de trabalho desenvolvida para avaliar a biblioteca BigDecimal do Java.
MPEGaudio	Aplica uma carga de trabalho com grande volume de operações em ponto flutuante, considerada adequada para avaliar performance na decodificação de arquivos MP3.
Scimark	Desenvolvida pelo NIST, essa carga de trabalho é amplamente difundida na indústria como *benchmark* de ponto flutuante. Uma de suas versões *(large)* é aplicada a um conjunto de dados de 32 MB e avalia o sistema de memória, enquanto o segundo *(small)* possui apenas 512KB e explora a JVM.
Serial	Avalia a serialiação de objetos e tipos primitivos, utilizando dados do *benchmark* JBoss.
Startup	Inicializa cada *benchmark* para execução de uma única operação, contabilizando o tempo de inicialização e finalização da JVM.
Sunflow	Executa uma carga de trabalho de visualização gráfica utilizando um algoritmo de iluminação global de código aberto.
XML	Executa operações do pacote *javax.xml.transform* sobre arquivos XML de tamanhos variados, de 1KB a 607KB.

3.3.1 SPECjvm2008

O SPECjvm2008 consiste em uma coleção de aplicações e *benchmarks* desenvolvida para testar a performance de um *Java Runtime Environment* (JRE), um ambiente utilizado para executar aplicações desenvolvidas em linguagem de programação Java. Os testes realizados medem o desempenho do JRE em um contexto específico, determinado pelo sistema operacional e o hardware nos quais o ambiente é executado [23].

A carga de trabalho do SPECjvm2008 é dividida em 11 *benchmarks* independentes, os quais contemplam operações comuns a diversas aplicações Java. Tais características refletem a intenção de tornar o SPECjvm2008 interessante para mensurar perfomance em uma ampla gama de sistemas que utilizem aplicações desenvolvidas nessa linguagem [38]. A Tabela 3.1 apresenta uma breve descrição de cada *benchmark* incluído na coleção SPECjvm2008.

Uma execução bem sucedida do SPECjvm2008 produz, para cada um dos 11 *benchmarks*, um resultado que reflete a frequência (em operações por minuto) na qual o sistema foi capaz de completar chamadas daquela carga de trabalho. Com isso, a principal saída do SPECjvm2008 consiste em 11 métricas de desempenho, uma para cada *benchmark* avaliado.

Quanto ao método de testes experimentais, cada máquina virtual executou uma rotina de 20 execuções em série do SPECjvm2008, armazenando os resultados obtidos em arquivos texto - ao final da execução da rotina, teriam sido produzidos 20 índices para cada um dos *benchmarks*. Em seguida, foram calculados o desvio padrão e a a média aritmética desses 20 índices, gerando assim 11 coeficientes de performance, um para cada *benchmark*, utilizados para comparação de resultados com outras instâncias.

Cada execução completa do SPECjvm2008 tem duração de pouco mais de duas horas, ou seja, os coeficientes de desempenho obtidos pela rotina desenvolvida neste trabalho fazem uma projeção com base no desempenho das 40 primeiras horas de execução, e assumem que essa projeção é válida para todo o tempo de vida da instância. Além disso, visando minimizar o número de variáveis do experimento, todas as intâncias de ambos provedores executaram o SPECjvm2008 utilizando o *JRE* 7, disponibilizado gratuitamente pela Oracle [39], em um sistema operacional Ubuntu Server 14.04 LTS.

3.3.2 Sysbench *fileio*

Caracterizado como um instrumento de *benchmark* multi-plataforma, o Sysbench [40] é uma ferramenta cuja proposta consiste em mensurar a performance de sistemas rapidamente, sem a necessidade de configurar ambientes complexos. A execução de um teste Sysbench resulta em um coeficiente de desempenho definido em operações por segundo, número esse composto pelo total de chamadas completas da carga de trabalho, dividido pelo tempo de duração do teste (em segundos). Assim, partindo dessa proposta, a ferramenta foi desenvolvida com 6 módulos de teste, de maneira que cada um deles executa diferentes cargas de trabalho:

- CPU: cada requisição consiste no cálculo de números primos até um valor especificado por um parâmetro, cálculo esse utilizando operações com inteiros de 64 bits;
- Threads: avalia o escalonador do sistema, explorando situações nas quais diversas *threads* competem pelo uso de recursos compartilhados (*mutexes*);
- Mutex: simula a situação na qual as *threads* executam de maneira concorrente durante a maior parte do tempo, adquirindo o controle do *mutex* por um curto intervalo de tempo;

- Memory: executa as operações sequenciais de leitura ou de escrita em memória;
- Fileio: produz diversas cargas de trabalho I/O, executando operações de leitura e/ou escrita em um domínio de disco (*workspace*) definido durante a configuração do teste;
- OLTP: avalia a performance de um banco de dados relacional, executando variadas operações de busca, criação, leitura, atualização e remoção de registros no banco.

Neste trabalho, apenas o módulo de teste *fileio* foi utilizado nos experimentos, de maneira que os resultados obtidos no *benchmark* representam o domínio de aplicações limitadas por operações de leitura e de escrita. O módulo de teste foi configurado para realização de operações de leitura e de escrita aleatórias em uma proporção 3:2, configuração padrão do Sysbench *fileio*, e que consequentemente atribui maior peso para performance de leitura.

Quanto ao método de testes experimentais, cada máquina virtual executou o Sysbench *fileio* continuamente durante um período de seis horas, produzindo assim um único coeficiente de operações/segundo para cada instância. Os resultados utilizados na análise sob perspectiva do Sysbench fazem uma projeção com base no desempenho dessas seis horas de execução, considerando que o coeficiente obtido representa o desempenho da instância durante toda a sua execução. Dessa forma, visando minimizar o número de variáveis do experimento, todas as intâncias de ambos os provedores executaram o Sysbench *fileio* em um sistema operacional Ubuntu Server 14.04 LTS.

Assim, após apresentar os provedores avaliados e o método utilizado para produção de resultados experimentais, a subseção a seguir apresenta alguns trabalhos da literatura que possuem características similares com a análise proposta neste trabalho.

3.4 Trabalhos Relacionados

Em análise da literatura, é possível constatar que métodos de comparação entre nuvens são propostos desde o início do paradigma. Garg *et al.* [41] propuseram, em 2010, um *framework* que comparasse diferentes provedores de nuvem com base em requerimentos definidos pelo usuário, e assim ajudá-lo a escolher o serviço que melhor se adequasse às suas necessidades. A performance dos serviços é monitorada durante seu uso e registrada para comparação com o esperado segundo suas SLAs. Desse modo, a tomada de decisão sobre o provedor recomendado é baseada em experiências prévias do usuário em vez de execução de *benchmarks*.

Quanto a medição de performance de máquinas virtuais, o uso de testes com *benchmarks* é um conhecido método de produção de resultados [21] [42]. Juve *et al.* [43]

avaliaram a performance e o custo registrados na execução de três diferentes aplicações científicas em nuvem Amazon, variando configurações de armazenamento em nuvem e a quantidade de recursos dedicados às cargas de trabalho. Tempo final (em horas) e custo (em dólares) das execuções foram as métricas consideradas para avaliar o custo-benefício das máquinas, de maneira similar à proposta deste trabalho. O trabalho em questão analisou apenas instâncias dedicadas oferecidas pela Amazon, sem considerar máquinas preemptáveis ou de outros provedores.

Hitoshi Oi [38] utilizou o *benchmark* SPECjvm2008 para avaliar o desempenho de três ambientes de hardware diferentes, executando JVMs de diferentes proprietários em cada um deles, e comparando os resultados obtidos nos testes do *benchmark*. O trabalho em questão traz o foco para o impacto de características como velocidade de *clock* do processador e hierarquias de *cache*.

No que se refere ao estudo de máquinas preemptáveis, diversos autores apresentaram propostas que agregassem confiabilidade às execuções. Yi *et al.* [32] avaliaram o uso de mecanismos de *checkpoint* para minimizar os problemas causados pela volatilidade das *Spot Instances* oferecidas pelo provedor Amazon, e ainda assim usufruir das vantagens dessas máquinas preemptáveis. Os autores concluíram que estratégias de *checkpoint* dinâmico reduziram significantemente custos monetários e aumentaram a confiabilidade dos recursos provisionados, mas seus testes se limitaram às instâncias preemptáveis de um único provedor.

Tatlow e Piccolo [44] compararam os custos entre um ambiente arquitetado em *cluster Kubernetes* e outro baseado em máquinas preemptáveis do provedor Google, executando cargas de trabalho relacionadas ao processamento de amostras de RNA sequenciado. Ainda que 11% das execuções tenham sido interrompidas, os resultados obtidos em instâncias preemptáveis apresentaram custo 49,2% menor que nos testes em *cluster*. Em suas conclusões, os autores sugerem o uso de instâncias preemptáveis como um método promissor para redução de custos, contanto que a ocorrência de preempções continue ocasional.

Diante do exposto, o capítulo a seguir apresenta os principais resultados obtidos nos testes experimentais e em simulações de cargas de trabalho baseadas em projeções desses testes.

Capítulo 4

Resultados

Tendo em vista a proposta apresentada, este capítulo busca identificar as instâncias preemptáveis dos provedores Google e Amazon com melhor custo-benefício nos *benchmarks* propostos, além de investigar eventuais vantagens comparativas de cada provedor. Para tal, este capítulo é dividido em 3 seções. A Seção 4.1 explora os resultados obtidos a partir do SPECjvm2008, a aplicação escolhida neste trabalho para representar aplicações de processamento intesivo. A Seção 4.2 comtempla os resultados obtidos a partir de execuções do *benchmark* Sysbench *fileio*, cuja carga de trabalho consiste em um fluxo intensivo de operações de leitura e escrita em disco. Por último, a Seção 4.3 aborda informações que não estão diretamente relacionadas à performance das instâncias, mas que podem ser relevantes para a análise comparativa proposta.

4.1 *Benchmarks* SPECjvm2008

A análise e apresentação de resultados referentes aos *benchmarks* SPECjvm2008 foi dividida da seguinte maneira: a Subseção 4.1.1 faz uma breve apresentação dos dados produzidos e descreve o comportamento das máquinas em relação às preempções; a Subseção 4.1.2 inspeciona os resultados obtidos com o objetivo de identificar o gargalo computacional dos *benchmarks* SPECjvm2008, informação essa de grande importância para uma escolha adequada de máquinas virtuais; a Subseção 4.1.3 investiga, com base em uma simulação, quais instâncias dos provedores Google e Amazon são as mais recomendadas para a execução dos *benchmarks* SPECjvm2008 com custo-benefício ótimo; a Subseção 4.1.4 explora características personalizáveis das máquinas virtuais preemptáveis e seu impacto no custo-benefício das execuções; por fim, a Subseção 4.1.5 realiza uma comparação objetivo entre as máquinas sugeridas entre os provedores Amazon e Google, avaliando a possível dominância de um deles.

Tabela 4.1: Amostras do *dataframe* produzido a partir do SPECjvm2008.

Instância	**Execuções**	**Benchmark**	**Média (ops/min)**	**Desvio (ops/min)**	**Preço (US$/hora)**
Google standard	30	compiler	58.6067	1.2931	0.010972
Google standard	30	compress	47.6503	0.625963	0.010972
Google dupla	21	crypto	67.3838	3.83339	0.031875
Google dupla	21	derby	107.153	7.65109	0.031875
Amazon t3small	20	mpegaudio	51.824	0.262895	0.0072375
Amazon t3small	20	scimark_large	28.6255	1.01945	0.0072375
Amazon t3small	20	scimark_small	111.452	2.99289	0.0110375
Amazon t3small	20	serial	62.304	2.97887	0.0065375
Amazon t3small	20	startup	25.3655	2.16018	0.0065375
Amazon c52xlarge	20	sunflow	123.882	1.075	0.1641375
Amazon c52xlarge	20	xml	820.991	4.83343	0.1641375

4.1.1 Resultados Iniciais

Como descrito no capítulo anterior, a execução de testes foi configurada por meio de uma rotina que realizasse 20 execuções em série dos *benchmarks* que compõem o SPECjvm2008. Assim, finalizados os testes, as amostras foram concentradas em um único *dataframe* e suas informações foram dispostas graficamente para uma interpretação mais intuitiva.

A Tabela 4.1 lista alguns exemplos de amostras inclusas neste *dataframe*, cujas colunas descrevem as características das máquinas e os resultados obtidos por elas. Os nomes atribuídos às instâncias deste capítulo advém de uma classificação atribuída pelos próprios provedores, incluindo eventuais traduções livres de acordo com a região da instância. As Figuras 4.1 a 4.2 apresentam - para Amazon e Google, respectivamente - as máquinas contempladas pelos testes SPECjvm2008, e o quantitativo de execuções em cada máquina.

Todas as máquinas virtuais configuradas no ambiente Amazon *Elastic Compute Cloud* (EC2) executaram a rotina sem interrupções ou preempções, e completaram exatamente as 20 execuções previstas, correspondendo ao comportamento esperado (vide Figura 4.1). Os testes realizados na *Google Compute Engine* (GCE), em contrapartida, revelam um

gráfico com o total de execuções altamente variável, com amostras apresentando entre 10 e 40 execuções (Figura 4.2). Tal comportamento se deu devido ao modelo de preempção praticado pelo provedor, no qual a interrupção pode ocorrer a qualquer momento entre as primeiras 24 horas de execução da instância [18].

Após cada episódio de preempção, a máquina e sua rotina foram reiniciadas, de maneira que as execuções previamente finalizadas fossem persistidas. Esse processo de *reboot* após cada preempção foi repetido em cada instância até que atingisse ao menos 10 execuções completas, ainda que não necessariamente contínuas. Uma vez que cada execução completa do SPECjvm2008 possui duração total de um pouco mais de 2 horas, todas as instâncias testadas em ambos os provedores tiveram execução total de pelo menos 20 horas [45]. Os campos *Média* e *Desvio* do *dataframe* dizem respeito ao desempenho médio do *benchmark* e ao desvio padrão registrado nas diversas execuções, respectivamente. Uma importante premissa deste capítulo considera o intervalo de tempo analisado como suficiente para que o valor médio observado possa ser projetado para todo o tempo de vida da instância.

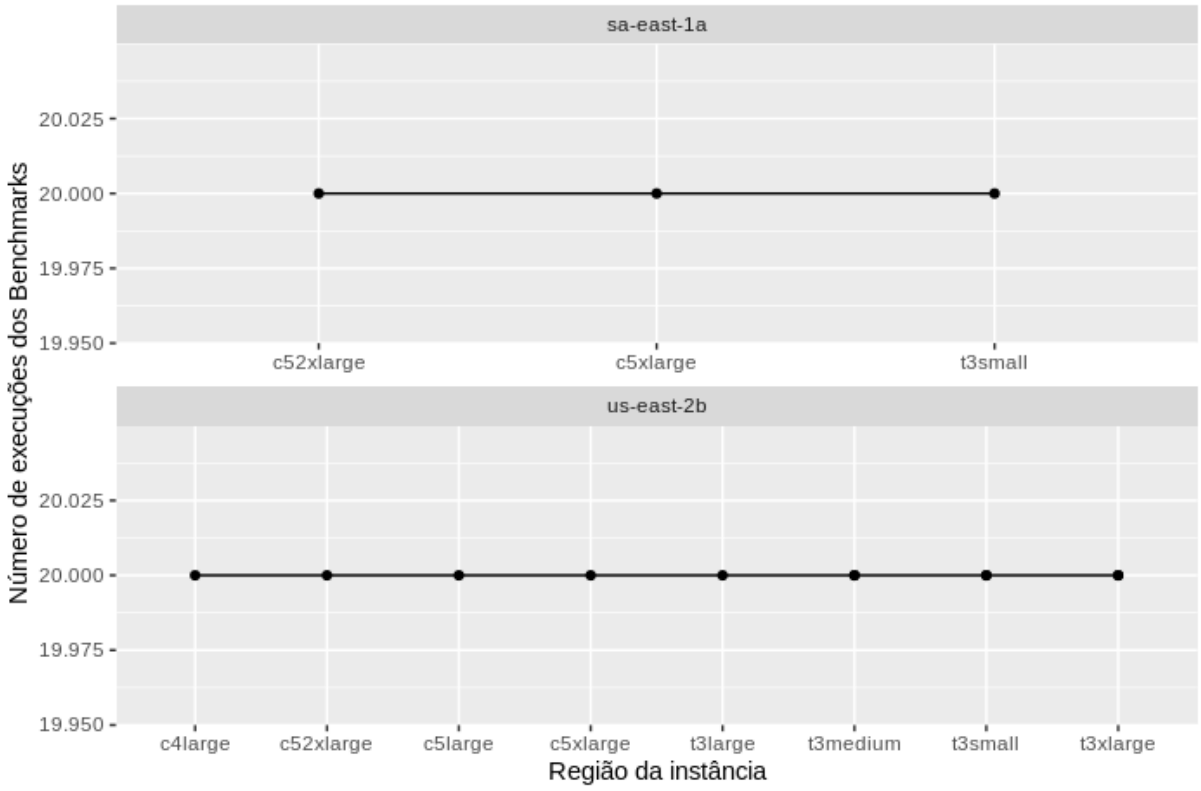

Figura 4.1: Quantidade de execuções para cada instância (Amazon).

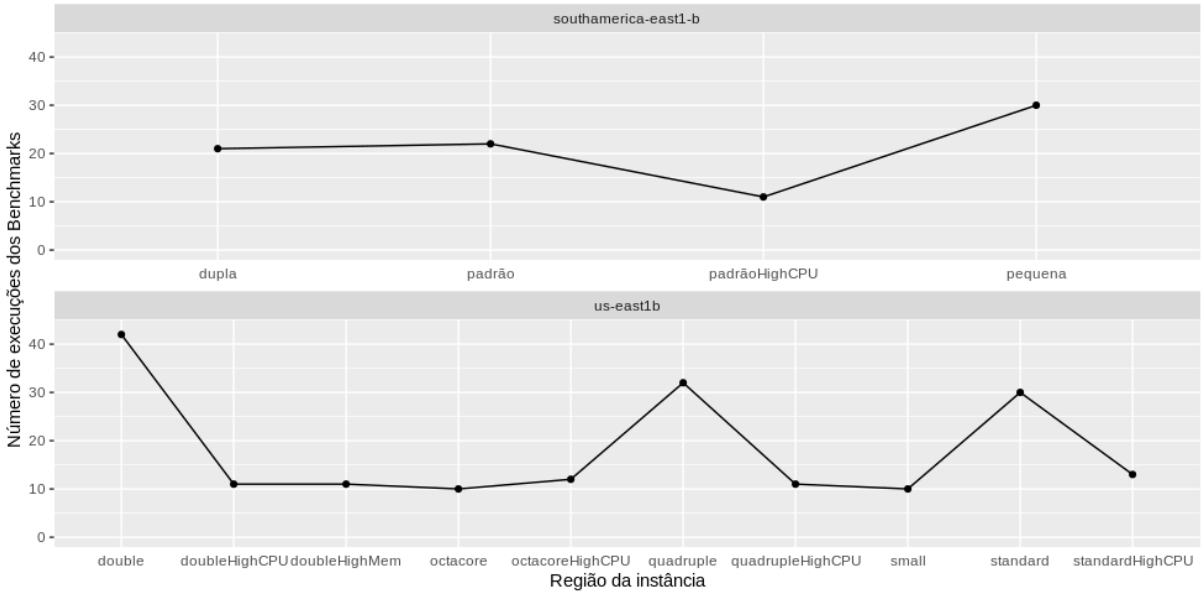

Figura 4.2: Quantidade de execuções para cada instância (Google).

Tabela 4.2: Máquinas básicas sugeridas pelos provedores.

Instância	**Provedor**	**vCPUs**	**RAM (GB)**
small	Google	0,5	1,7
standard	Google	1	3,75
double	Google	2	7,5
quadruple	Google	4	15
octacore	Google	8	30
c5large	Amazon	2	4
c5xlarge	Amazon	4	8
c52xlarge	Amazon	8	16

4.1.2 Gargalo Computacional

Esta subseção busca, a partir de resultados obtidos experimentalmente, identificar o gargalo de desemepenho do SPECjvm2008. Como explicado no capítulo anterior, cada instância preemptável avaliada por este trabalho produziu 11 coeficientes SPEC, medidos em operações por minuto e referentes aos 11 *benchmarks* da aplicação. A Tabela 4.2 apresenta as configurações de hardware descritas para as instâncias padronizadas oferecidas pela Google, e para as instâncias da família C5Large da Amazon. Visando reduzir o número de variáveis da análise, apenas amostras de regiões *us-east*, localizadas no Leste dos Estados Unidos, foram consideradas nestes gráficos. Além disso, este trabalho contemplou prioritariamente máquinas de pequeno porte, com até 8 vCPUs.

As Figuras 4.3 a 4.4 apresentam os resultados obtidos pelas instâncias preemptáveis dos

provedores Google e Amazon, respectivamente. O *eixo X* representa o custo em dólares por cada hora de execução da instância, e o *eixo Y* retrata o número médio de operações executadas por minuto. Uma vez que cada *benchmark* possui seu coeficiente, a figura apresenta o desempenho de cada um deles separadamente. Além disso, foram aplicados modelos lineares sobre as amostras, representados pelas linhas em vermelho. Em leitura dos gráficos é possível perceber que apenas o *benchmark startup* apresentou amostras destoantes do modelo linear, comportamento este justificado por sua carga de trabalho limitada por outros fatores que não processamento intensivo [46]. Em contrapartida, os demais apresentaram resultados bem comportados linearmente em ambos os provedores, ou seja, instâncias mais caras sugerem um melhor desemepenho na execução.

Uma vez que as instâncias de custo mais elevado consistem naquelas que possuem mais recursos computacionais, é possível supor inicialmente que a adição de recursos computacionais implica em melhor performance na execução dos *benchmarks*. Esta subseção busca investigar, com base em testes práticos, quais recursos trazem maior ganho de desempenho para os *benchmarks* que compõem o SPECjvm2008.

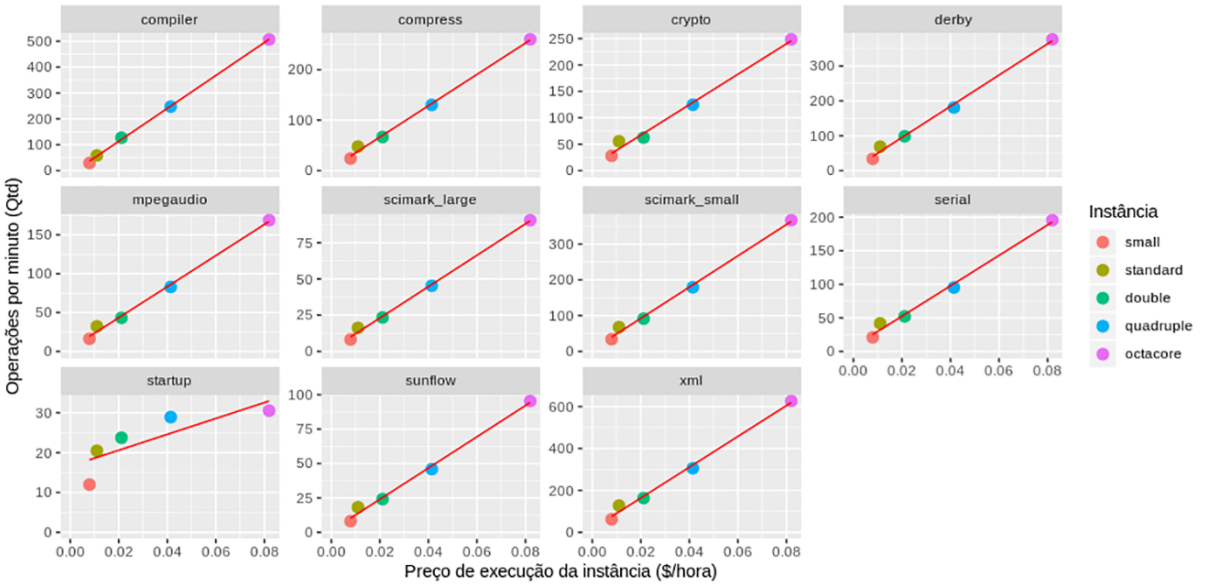

Figura 4.3: Desempenho SPECjvm2008 por *benchmark* (Google).

Assim sendo, com o objetivo de avaliar o impacto da memória disponível na performance das máquinas virtuais, foram executados testes personalizando a quantidade de RAM alocada nas instâncias Google, utilizando como referência as configurações padronizadas utilizadas anteriormente. Com isso, a nomenclatura das novas instâncias recebeu um sufixo: *HighMem* caso sua memória seja comparativamente maior do que o valor original, e *HighCPU* caso sua memória seja menor. As configurações de instâncias Google contempladas pela análise do SPECjvm2008 são apresentadas na Tabela 4.3.

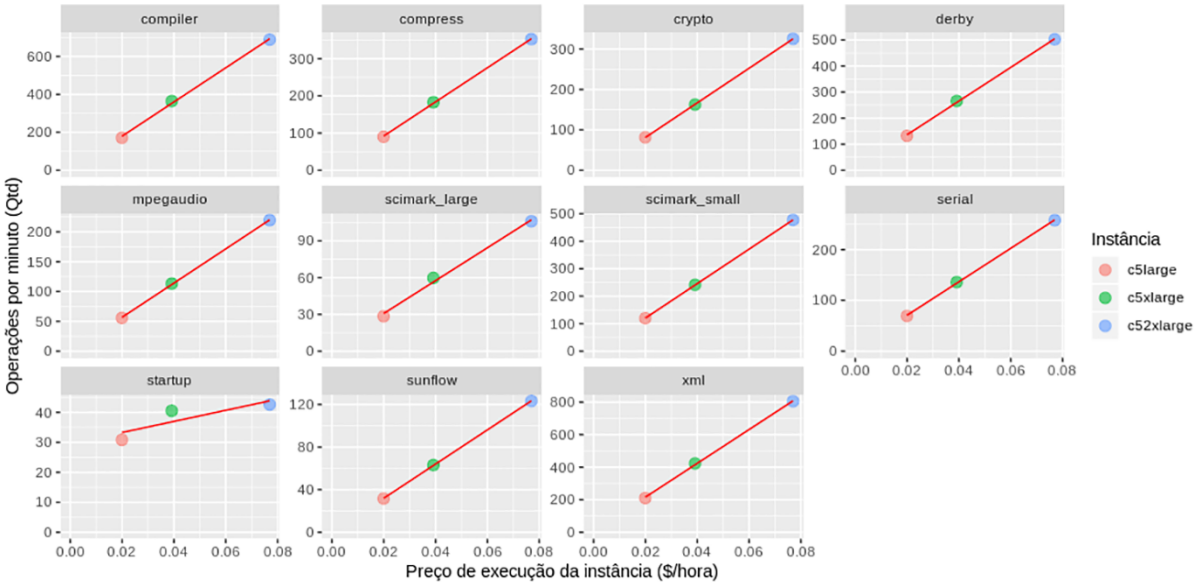

Figura 4.4: Desempenho SPECjvm2008 por *benchmark* (Amazon).

Tabela 4.3: Máquinas Google avaliadas.

Instância	**vCPUs**	**RAM (GB)**
standardHighCPU	1	1,5
standard	1	3,75
doubleHighCPU	2	1,8
double	2	7,5
doubleHighMem	2	13
quadrupleHighCPU	3,6	22,5
quadruple	4	15
octacoreHighCPU	8	7,2
octacore	8	30

A Figura 4.5 apresenta as amostras obtidas nas instâncias padronizadas em conjunto com as execuções personalizadas. Comparando os resultados obtidos no *eixo Y* do gráfico (Figura 4.5) é possível observar que instâncias com diferentes configurações de RAM apresentaram coeficientes de operações por minuto similares entre si, desde que fossem configuradas com o mesmo número de vCPUs. A partir dessa comparação, infere-se que o desempenho dos *benchmarks* SPECjvm2008 é diretamente impactado pelo número de vCPUs da máquina virtual, e pouco sensível à memória RAM disponível.

Uma vez que a alocação de memória adicional aumenta consideravelmente o custo de execução da instância (vide *eixo X* da Figura 4.5) e não apresenta ganho de performance, a categoria de instâncias *HighCPU* se revela como a mais eficiente economicamente para execução dos *benchmarks* SPECjvm2008 dentre as instâncias Google.

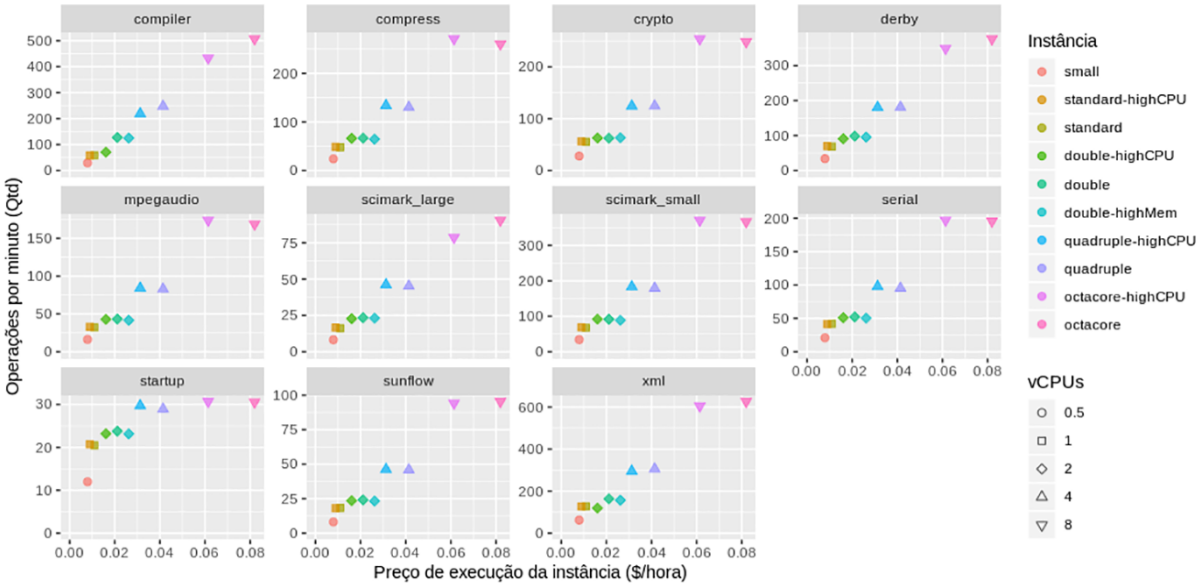

Figura 4.5: Desempenho SPECjvm2008 por *benchmark* incluindo instâncias *High* (Google).

Tabela 4.4: Máquinas Amazon avaliadas.

Instância	**vCPUs**	**RAM (GB)**
t3small	2	2
t3medium	2	4
t3large	2	8
t3xlarge	4	16
c4large	2	3,75
c5large	2	4
c5xlarge	4	8
c52xlarge	8	16

A análise do impacto da memória RAM no desempenho das instâncias foi realizada de maneira diferente nas instâncias Amazon, uma vez que esse provedor não disponibiliza o mesmo nível de customização na alocação de seus recursos computacionais. Assim, para este teste foram utilizadas máquinas de diferentes famílias (T3, C4, C5), cujas configurações são listadas na Tabela 4.4.

A Figura 4.6 apresenta o desempenho relatado por cada amostra (em operações por minuto) em função da memória RAM alocada para a instância em questão. As amostras sugerem que a memória RAM disponível *(eixo X)* tem baixa correlação com o coeficiente de operações por minuto da instância *(eixo Y)*. Como exemplo, é possível ver que todas as máquinas configuradas com 2 vCPUs (em vermelho) apresentaram performance similar nos 11 *benchmarks*, mesmo variando a RAM entre 2 GB e 8 GB. Este comportamento

reafirma a hipótese proposta de que a adição de memória RAM não implica em ganho de performance relevante na execução do SPECjvm2008. Com isso, conclui-se que o recurso computacional limitador do desempenho desses *benchmarks*, conhecido como gargalo [47], se encontra no número de vCPUs disponíveis.

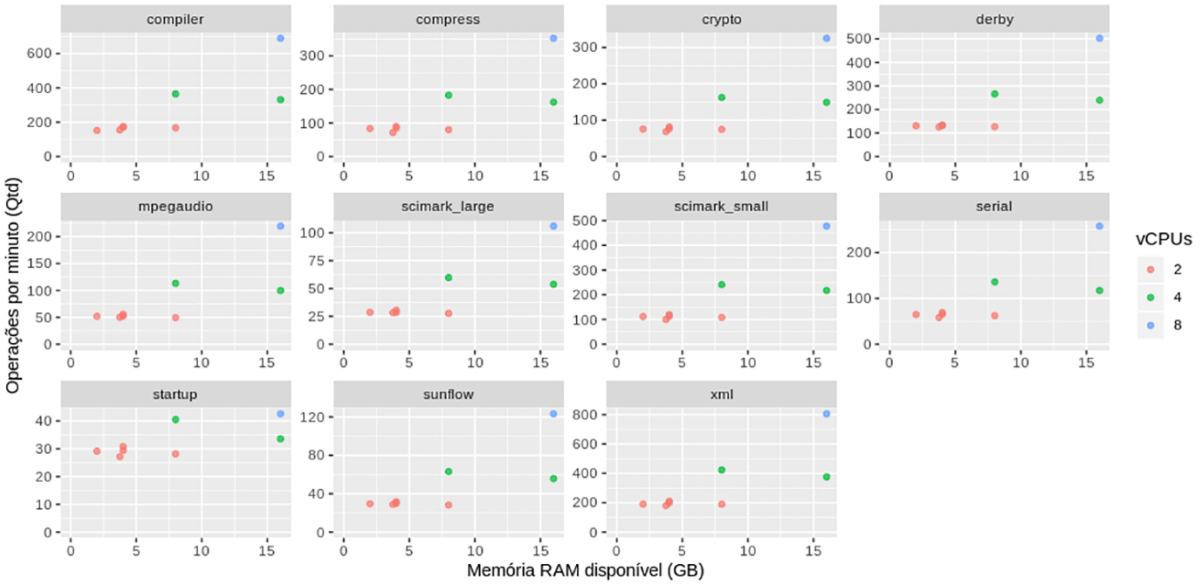

Figura 4.6: Desempenho SPECjvm2008 por número de vCPUs (Amazon).

4.1.3 Instâncias Recomendadas

Para um usuário que executa cargas de trabalho em nuvens públicas, custo mínimo e tempo de execução baixo são características de grande relevância. Esta subseção busca identificar as instâncias dos provedores Google e Amazon que oferecem menor custo e tempo de resposta na execução do *benchmark* SPECjvm2008. Para isso, uma das maneiras mais simples de comparar o custo-benefício oferecido por diferentes instâncias é apresentando resultados quantitativos de execuções reais e/ou simuladas.

Tendo isso em vista, os gráficos desta subseção são baseados na simulação de uma carga de trabalho composta por um milhão de operações do *benchmark* avaliado. O *eixo X* dos gráficos apresenta o tempo de execução da carga em horas, enquanto o *eixo Y* descreve o custo em dólares para a execução da mesma. Com essa apresentação, é possível contrastar preços e performances das máquinas preemptáveis mais facilmente.

A Figura 4.7 descreve a performance das instâncias Google na simulação proposta. Em leitura do gráfico (Figura 4.7), a instância *small* apresenta uma nítida desvantagem em relação às demais máquinas, com tempo de execução muito superior em todos os *benchmarks*, e custo mais elevado em grande parte dos gráficos apresentados. Essa confi-

guração caracteriza máquinas com vCPU compartilhada, de maneira que o recurso físico é compartilhado entre diversas instâncias *small*. Ainda que esta seja a instância com menor custo por hora de execução (vide Figura 4.5), o extenso período para execução da carga de trabalho torna a simulação na máquina *small* tão dispendiosa quanto em máquinas de maior poder computacional.

Além dessa constatação, é possível perceber que nenhuma das instâncias Google contempladas pela análise apresenta custo-benefício universalmente dominante. Em uma avaliação baseada exclusivamente no custo, a instância *standardHighCPU* se destaca como a mais recomendada uma vez que apresentou valor monetário mais baixo em 10 dos 11 *benchmarks*. Em contrapartida, as demais instâncias *HighCPU* executaram a carga de trabalho em um tempo total consideravelmente menor, o que pode justificar o aumento de custo para uma parcela dos usuários. Caso a avaliação de eficiência seja feita por um viés de tempo total de execução, a instância *octacoreHighCPU* seria vista como a melhor dentre as máquinas Google avaliadas. Uma vez que grande parte dos usuários de serviços de nuvem buscam um equilíbrio entre as duas métricas, é importante que o cliente pondere suas prioridades para que execute instâncias preemptáveis mais adequadas ao seu perfil.

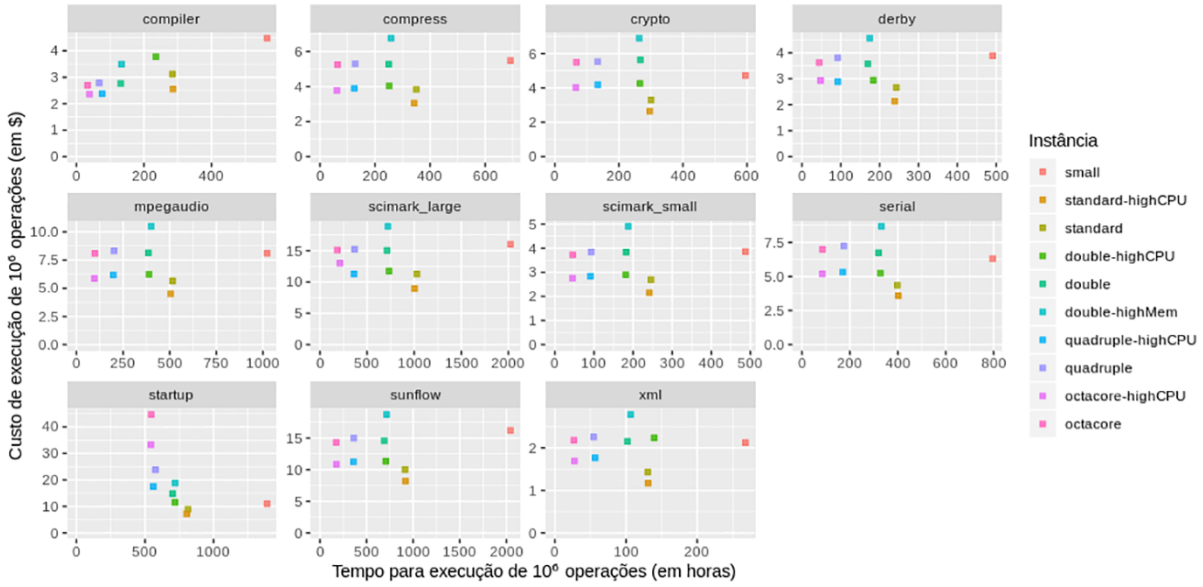

Figura 4.7: Execução de uma carga de trabalho SPECjvm2008 com um milhão de operações (Google).

Constatado que poder de processamento é o principal gargalo do SPECjvm2008, foram executadas as instâncias que apresentassem os maiores índices de vCPU por RAM entre as disponibilizadas pela Amazon, os quais são compostos pela divisão do número de vCPUs pela memória RAM disponível (em GB). Essa métrica foi adotada visando minimizar o dispêndio monetário com memória subutilizada, identificada na subseção anterior. Os

resultados obtidos pelas instâncias Amazon na simulação proposta são apresentados na Figura 4.8, cujos testes executados envolveram configurações de diversas famílias.

Em leitura do gráfico (Figura 4.8), é possível observar que as execuções mais econômicas foram aquelas realizadas em máquinas virtuais com maiores índices de vCPU por RAM, como previsto anteriormente. Ainda assim, as execuções da família C4 apresentaram desvantagem econômica e performática em relação a máquinas de coeficiente inferior. Segundo a Amazon [48], a família C5 é caracterizada por máquinas virtuais com processadores *Intel Skylake* superiores aos *Intel Haswell* da família C4, trazendo uma performance e custo melhores que a geração anterior. Dito isso, constata-se a importância de usuários de nuvem estarem atentos a eventuais evoluções e lançamentos de novas famílias de instâncias Amazon, uma vez que as migrações para instâncias mais novas devem, neste provedor, ser feitas pelo usuário.

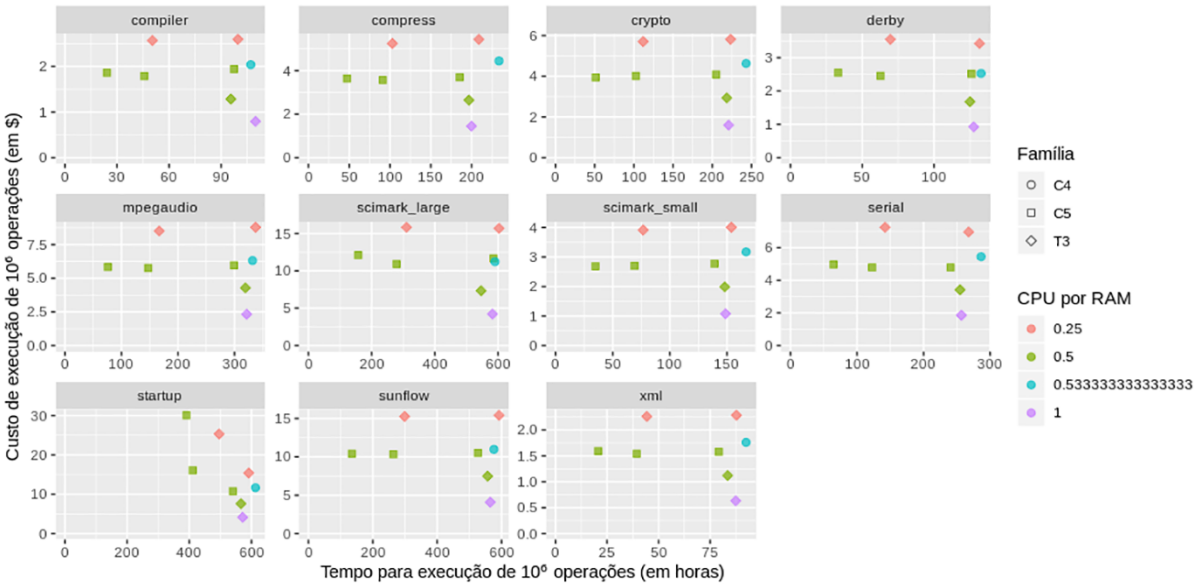

Figura 4.8: Execução de uma carga de trabalho SPECjvm2008 com um milhão de operações (Amazon).

Assim, assumindo as conclusões descritas nesta seção, a Figura 4.9 reapresenta as amostras da Figura 4.8 descartando aquelas de gerações anteriores ou que registraram um índice vCPU por RAM menor que 0,5 - restando assim apenas instâncias candidatas para melhor custo-benefício na execução do SPECjvm2008.

Comparando os pontos do gráfico (Figura 4.9), é possível constatar que não há uma categoria de máquinas preemptáveis que apresente liderança absoluta de desempenho. Assim como observado entre as máquinas Google, a instância Amazon que possui custo mais baixo é diferente daquela que apresenta tempo de execução mínimo, indicando um *tradeoff* entre as duas métricas. Sob uma perspectiva financeira, a configuração *t3small*

se mostra a mais econômica, enquanto a *c52xlarge* desponta como a mais eficiente em tempo total de execução.

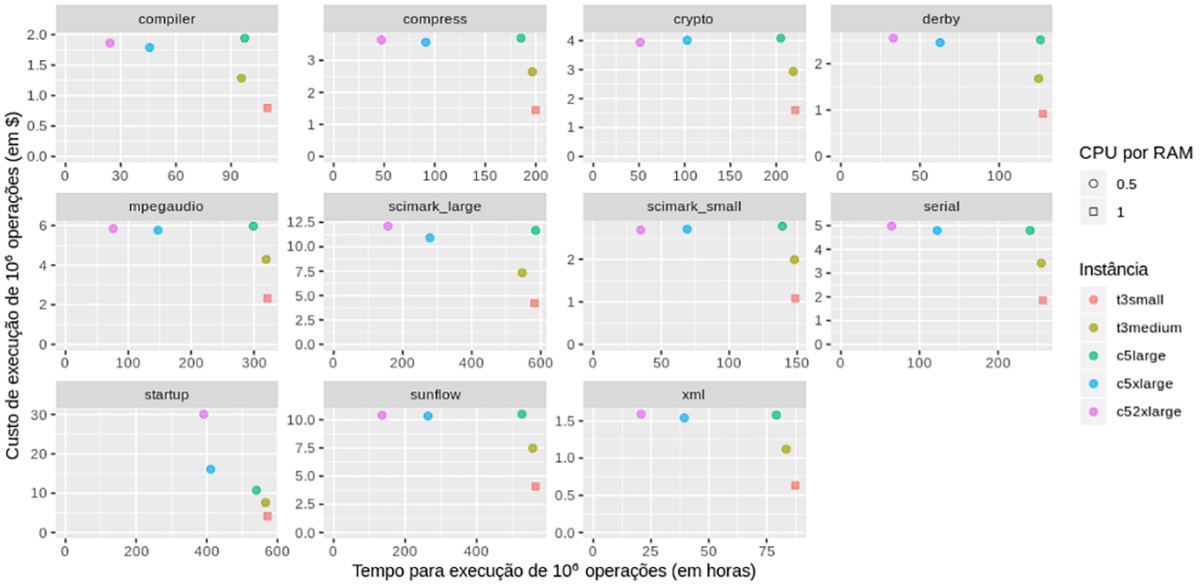

Figura 4.9: Execução de uma carga de trabalho SPECjvm2008 com um milhão de operações (instâncias Amazon selecionadas).

4.1.4 Impactos de Região e Fabricante de *hardware*

Tipicamente, os principais provedores de nuvem do mercado possuem *datacenters* em diversos lugares do Mundo, denominando cada um desses como uma região de nuvem disponível para seus usuários [49]. Uma vez que o provedor se encontre distribuído geometricamente, é possível que ele atenda a requisitos como latência máxima, segurança ou legislação de usuários em potencial, ampliando assim o seu mercado consumidor.

Usualmente, a região escolhida possui grande impacto no preço final da execução da instância preemptável criada. O preço praticado por um provedor em uma região se dá, entre outros fatores, pelo custo de operações na área onde o *datacenter* se encontra, pela oferta e demanda na região, e pela competitividade com outros provedores de nuvem próximos. Portanto, a escolha de regiões adequadas para a execução das instâncias pode tornar o serviço menos custoso para o usuário - principalmente em cargas de trabalho onde a localização geográfica da máquina virtual tem pouco impacto em seu desempenho [50]. Investigando ainda outras maneiras de reduzir custos para a carga de trabalho simulada, identificou-se que a Amazon oferece variações de *hardware* em algumas categorias de máquinas, variações essas com impacto direto no custo das instâncias. Diante desse

Tabela 4.5: Custos para execução de máquinas preemtáveis segundo região.

Instância	**Preço em região sugerida (US$/hora)**	**Preço em região local (US$/hora)**	**Aumento de preço (em %)**
small	0,0079	0,0118	49,4%
standard	0,0110	0,0166	50,1%
standardHighCPU	0,00893	0,01347	50,8%
double	0,0211	0,0319	50,2%
t3small	0,0063	0.0101	60,3%
c5xlarge	0.0382	0.0878	129,8%
c52xlarge	0.0760	0.1632	114,7%

cenário, esta subseção busca avaliar experimentalmente o impacto da escolha de regiões e de fabricante de *hardware* no desempenho dos *benchmarks* SPECjvm2008.

Primeiramente, algumas configurações de instâncias de ambos os provedores foram selecionadas para a execução de testes, variando apenas a região entre elas. Cada provedor apresenta uma região sugerida para a execução de máquinas virtuais na qual os custos de execução são os mais baixos dentre as regiões do provedor, o que motivou a adição das mesmas ao escopo deste trabalho. Para a Amazon, a região apresentada foi a *us-east-2b*, localizada em Ohio-USA, enquanto a Google Cloud sugeriu a região *us-east1b*, localizada na Carolina do Sul-USA [49] [51].

Além disso, os provedores Google e Amazon disponibilizam regiões de nuvem situadas na América do Sul, sendo estas denominadas *southamerica-east1* pela Google Cloud e *sa-east-1a* pela Amazon. Ambos os *datacenters* estão localizados em São Paulo-Brasil e suas instâncias foram contempladas pela análise deste trabalho devido à proximidade geográfica entre as regiões e o usuário, este último localizado no Distrito Federal-Brasil. Os preços e configurações das instâncias executadas nas regiões aqui citadas são apresentados na Tabela 4.5.

A análise de desempenho de instâncias de diferentes regiões foi realizada de maneira similar à Subseção 4.2.3, baseada na simulação de uma carga de trabalho com um milhão de operações do *benchmark* avaliado. As Figuras 4.10 a 4.11 apresentam as amostras obtidas pelos provedores Google e Amazon, respectivamente.

Analisando os gráficos das máquinas Google, as instâncias localizadas na América do Sul apresentaram custo de execução consideravelmente maior para todos os *benchmarks* nas diversas configurações de *hardware* amostradas, comportamento já esperado segundo análise da literatura [52]. Em contrapartida, essas mesmas instâncias locais registraram um tempo de execução menor em boa parte das amostras registradas. Com isso, constata-se que existem cenários onde o aumento de custo trazido pela adesão de nuvens em regiões Google mais dispendiosas pode ser compensado pelo aumento de performance da carga

de trabalho (vide *benchmarks* compiler e scimark_large).

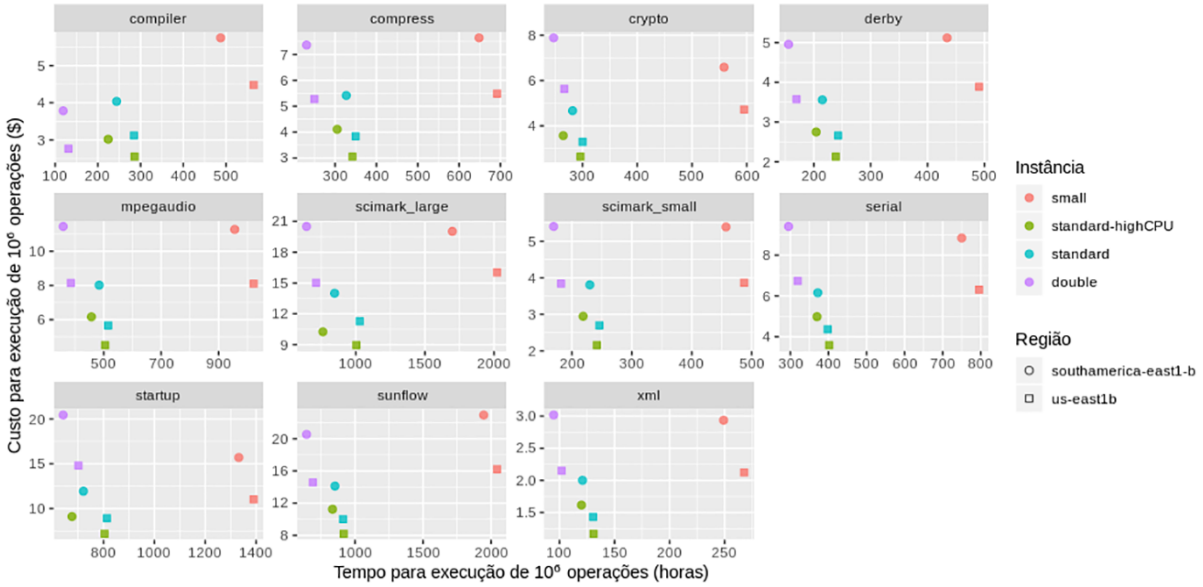

Figura 4.10: Custo e performance dos *benchmarks* SPECjvm2008 em diferentes regiões (Google).

Analisando os resultados em instâncias Amazon, pouca variação no tempo de resposta foi obtida na comparação entre uma máquina virtual localizada em Ohio-USA e uma instância equivalente localizada em São Paulo (vide Figura 4.11). Com isso, o preço adicional cobrado por hora de execução das máquinas virtuais brasileiras escala diretamente no preço final, revelando a superioridade econômica da região de Ohio para fins de execução dos *benchmarks* SPECjvm2008.

Dessa maneira, buscando outros métodos de reduzir custos para a carga de trabalho simulada, identificou-se que a Amazon oferece uma configuração, ainda não explorada neste trabalho, na qual é possível definir que as instâncias sejam executadas em hardware com processadores produzidos pela AMD, enquanto as máquinas padronizadas utilizam dispositivos Intel.

Levantada a suposição de que o SPECjvm2008 tenha custo e/ou tempo de execução reduzidos na execução de instâncias AMD, foram coletadas amostras de algumas máquinas virtuais preemptáveis provisionadas em hardware deste fabricante. As instâncias utilizadas nestes testes foram executadas na região *us-east-2b*, localizada em Ohio e sugerida pela Amazon. A Tabela 4.6 descreve as variações nos preços das instâncias, enquanto a Figura 4.12 apresenta graficamente o desempenho registrado pelas mesmas com a carga de trabalho simulada.

Analisando os resultados, é possível identificar que as máquinas Intel apresentaram tempo de execução menor, e custo equivalente nos *benchmarks compiler* e *scimark_large*,

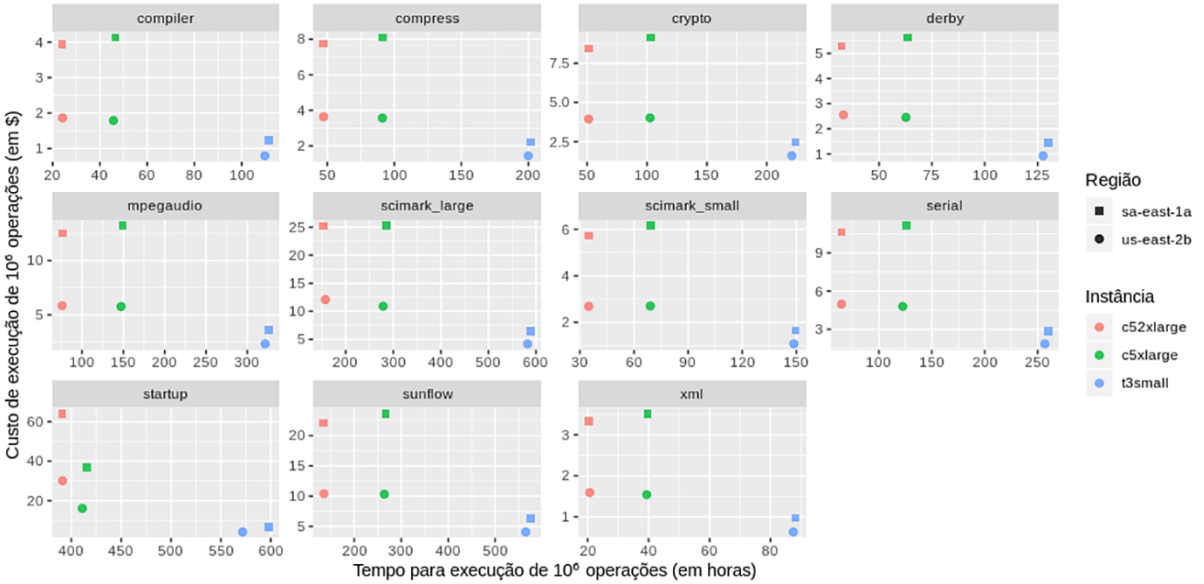

Figura 4.11: Custo e performance dos *benchmarks* SPECjvm2008 em diferentes regiões (Amazon).

Tabela 4.6: Preços de instâncias Amazon em diferentes fabricantes.

Instância	**Fabricante das CPUs**	**VMPrice AMD (US$/hora)**	**VMPrice Intel (US$/hora)**	**Aumento de custo com Intel (em %)**
t3small	AMD	0.0056	0.0063	12,5
t3medium	AMD	0.0113	0.0125	10,6
t3xlarge	AMD	0.0451	0.0501	11,1

enquanto as instâncias executadas em hardware AMD relataram menor tempo e preço total na execução *derby*. Além disso, há registros nos quais a execução AMD foi mais econômica em uma das configurações de instâncias e mais dispendiosa em outra (*sunflow* e *xml*, por exemplo). Com isso, conclui-se que as características da carga de trabalho e a categoria da máquina virtual instanciada são importantes informações para a determinação do fabricante mais indicado para a execução.

4.1.5 Comparação entre os Provedores

Vistos os resultados obtidos em cada provedor, outra característica relevante para a escolha de um serviço é a estabilidade de performance obtida nas instâncias oferecidas, seja no EC2 da Amazon ou no GCloud Compute Engine da Google. A Figura 4.13 busca explorar essa característica experimentalmente. Retomando o *dataframe* apresentado na Tabela 4.1, cada amostra produziu um valor de desvio padrão, além do valor médio explorado anteriormente. A Figura 4.13 apresenta, em porcentagem, o coeficiente de variação

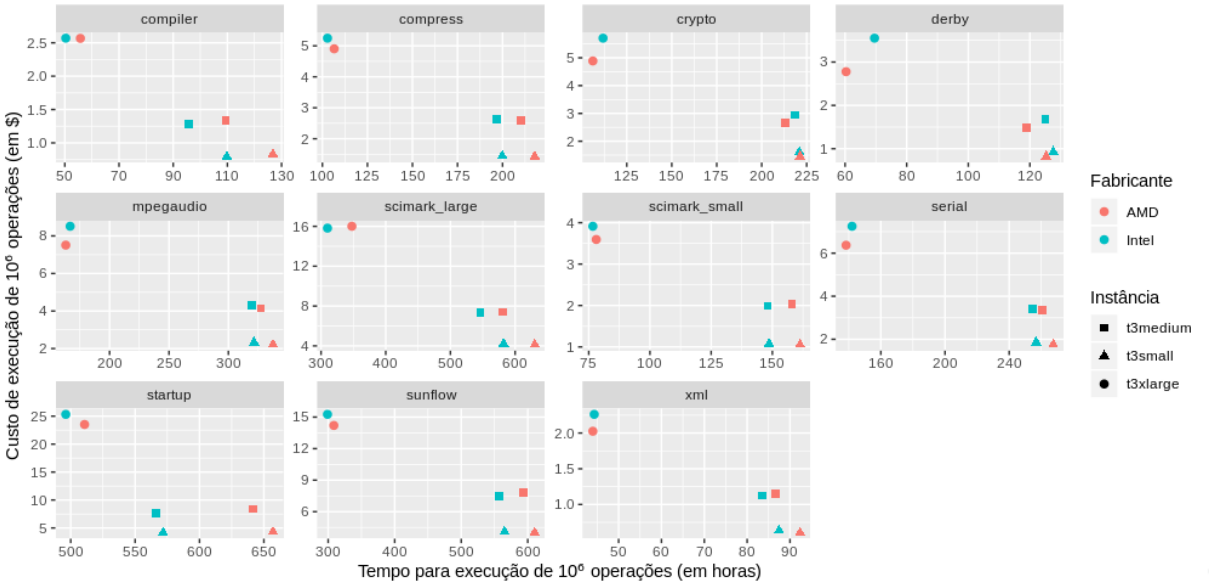

Figura 4.12: Desepenho SPECjvm2008 de diferentes fabricantes (Amazon).

obtido por cada amostra, o qual é definido como a razão do desvio padrão pela média [53]. Este indicador foi adotado uma vez que cada *benchmark* possui valores de média muito particulares, facilitando a visualização desejada.

Em leitura do gráfico, constata-se que as instâncias Google localizadas na região southamerica-east1-b registraram as maiores variações em grande parte dos *benchmarks*, o que sugere um desempenho mais instável entre as máquinas virtuais dessa região. Analisando as amostras das regiões *us-east-2b* e *us-east1b*, consideradas pelos provedores como regiões recomendadas, Amazon e Google apresentaram comportamento similar em grande parte dos *benchmarks* executados, ou seja, não sugerem uma recomendação de provedor.

Comentadas as principais características das máquinas virtuais preemptáveis nos provedores Amazon e Google, a Figura 4.14 apresenta as amostras que apresentaram melhores preços e/ou tempo de resposta entre os testes realizados nesta seção. Sob uma perspectiva que priorize o tempo de resposta, a instância *octacoreHighCPU* foi a que apresentou melhor performance dentre as máquinas Google, mas a instância Amazon *c52xlarge* registrou menor tempo de execução em todos os *benchmarks*, mantendo preços similares. Avaliando máquinas de categorias mais econômicas, a Amazon apresenta resultados ainda mais convidativos: suas instâncias *t3small* executaram a carga de trabalho em menos tempo e com economia de aproximadamente 40% em relação à instância Google mais econômica identificada (*standardHighCPU*). Conclui-se, com isso, que as instâncias preemptáveis Amazon EC2 apresentam melhor custo-benefício na execução de cargas de trabalho do *benchmark* SPECjvm2008 ou de aplicações semelhantes.

Concluída a disposição dos resultados obtidos a partir de testes com os *benchmarks*

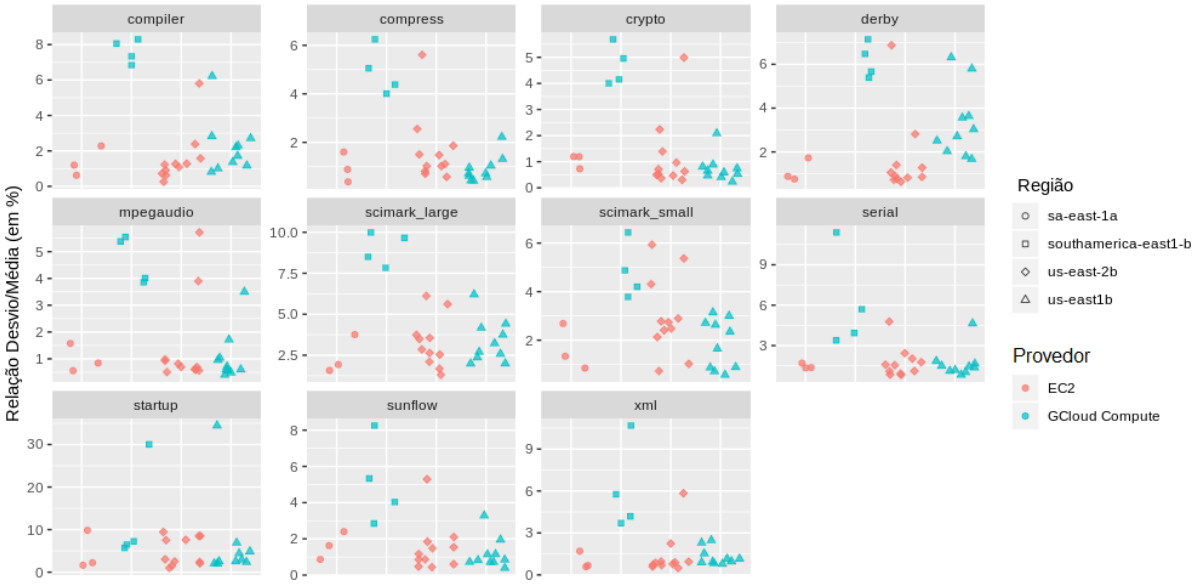

Figura 4.13: Coeficiente de Variação SPECjvm2008 (em %).

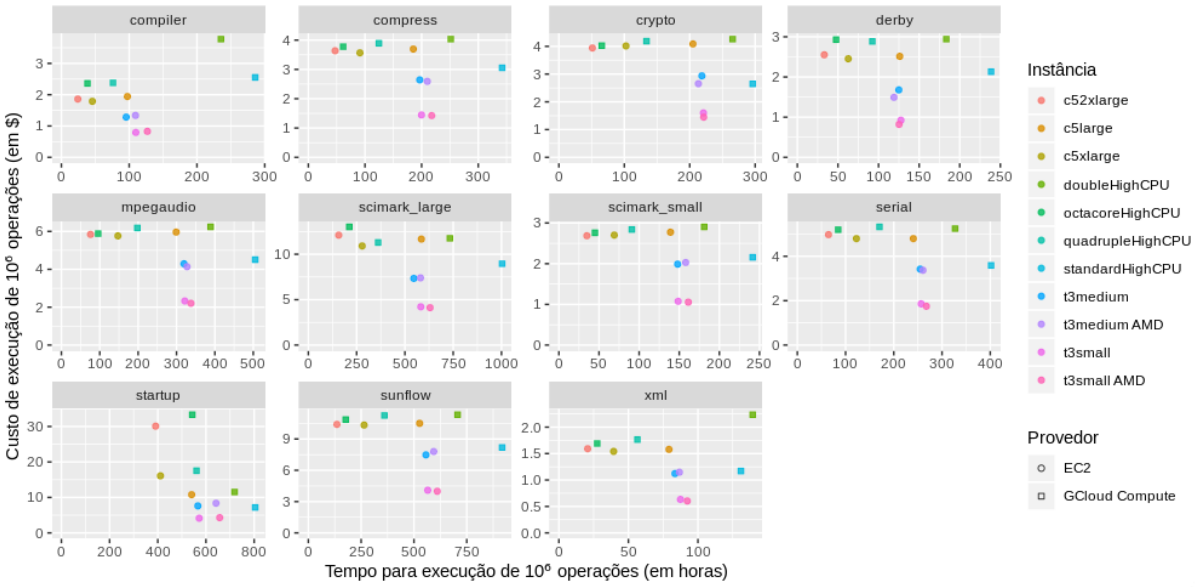

Figura 4.14: Execuções SPECjvm2008 nos provedores Google e Amazon.

SPECjvm2008, a seção a seguir apresenta uma análise a partir do *benchmark* Sysbench *fileio* em máquinas virtuais dos provedores Amazon e Google.

4.2 *Benchmark* Sysbench *fileio*

A análise e apresentação de resultados referentes ao *benchmark* Sysbench *fileio* foi dividida da seguinte maneira: a Subseção 4.1.1 faz uma breve apresentação dos dados produzidos e descreve o comportamento das máquinas em relação à preempções; a Subseção 4.1.2 inspeciona os resultados obtidos com o objetivo de identificar o gargalo computacional de uma carga de trabalho do *benchmark* SPECjvm2008, critério importante para atingir melhor performance na execução; a Subseção 4.1.3 investiga, com base em uma simulação, quais instâncias dos provedores Google e Amazon são as mais recomendadas para a execução do *benchmark* SPECjvm2008 com custo-benefício ótimo; a Subseção 4.1.4 explora características personalizáveis de máquinas virtuais preemptáveis e seu impacto no custo-benefício das execuções; por fim, a Subseção 4.1.5 realiza uma comparação objetivo entre as máquinas sugeridas entre os provedores Amazon e Google, avaliando a possível dominância de um deles.

4.2.1 Resultados iniciais

A Tabela 4.7 apresenta as principais informações sobre as máquinas contempladas pelos testes Sysbench *fileio*, cada uma delas contando com discos rígidos (HDD) de 100 GB e configuradas com um domínio de 80 GBs para leitura e escrita aleatórias de dados - domínio esse escolhido para que caracterizasse um volume de dados muito superior à memória RAM disponível nas instâncias. As máquinas Google e Amazon avaliadas nestes testes foram instanciadas nas regiões localizada em Carolina do Sul-USA e Ohio-USA, respectivamente. Tais regiões foram escolhidas pela vantagem econômica atribuída às mesmas segundo resultados obtidos nos testes SPECjvm2008.

Em leitura dos resultados apresentados na Tabela 4.7 fica evidente uma diferença entre os dois provedores de grande relevância: há uma cobrança feita pela Amazon devido a operações de leitura e escrita de dados (operações I/O) em suas unidades de disco, enquanto a Google inclui esse serviço à adesão de suas máquinas virtuais. Para uma aplicação de grande volume de operações dessa categoria, essa despesa é de grande relevância na identificação do melhor custo-benefício. Essa característica sugere que instâncias preemptivas Amazon são mais interessantes para aplicações de alto processamento, como o *benchmark* SPECjvm2008, que para aplicações caracterizadas por um intenso fluxo de operações I/O, como é o caso do *benchmark* Sysbench *fileio*.

Enquanto o SPECjvm2008 conta com 11 processos de trabalho singulares e produz resultados individuais para cada um deles, a execução do Sysbench em modo *fileio* produz um único coeficiente para análise e comparações. Diferentemente do SPECjvm2008, as instâncias escolhidas para a execução dos testes *fileio* possuem quantidades reduzidas de

Tabela 4.7: Execuções Sysbench *fileio* de 80 GB em dispositivos HDD padrão.

Instância	**vCPU**	**RAM (GB)**	**VMPrice (US$/hora)**	**IOPrice (US$/hora)**	**DiskPrice (US$/hora)**	**Média (Ops/Seg)**
Google t3nano	2	0.5	0.0016	0.0444186	0.00694	246.77
Google t3small	2	2	0.0063	0.0454734	0.00694	252.63
Google t3medium	2	4	0.0125	0.044667	0.00694	248.15
Amazon standard	1	3.75	0.01014	0	0.00556	86.43
Amazon standard HighMem	1	6.5	0.012625	0	0.00556	89.24
Amazon small06	0.5	0.6	0.003556	0	0.00556	85.26

vCPUs uma vez que essa carga de trabalho é caracterizada por pequenas demandas de processamento [40].

A Figura 4.15 apresenta os resultados obtidos no teste em questão, separando as amostras de acordo com o seu provedor. No *eixo X* está representado o custo total de execução do *benchmark*, composto pelo custo de execução da instância somado ao custo de manutenção do disco e de operações de leitura e escrita. O *eixo Y* retrata a média de requisições de leitura e/ou escrita por segundo ao longo do período de execução do Sysbench, que neste caso foi de 6 horas. As legendas indicam o número de vCPUs e a quantidade de memória RAM (em GB) disponibilizados para a instância, sendo que a amostra com 0.5 vCPU caracteriza a máquina virtual Google com compartilhamento de *hardware* apresentada na Subseção 4.1.3.

Em leitura dos resultados, é possível observar pelos gráficos (Figura 4.15) que o aumento de memória RAM disponível tem pouco impacto no desemepenho do *benchmark*, de maneira similar ao teste realizado sob perspectiva do SPECjvm2008. Além disso, é possível concluir que um aumento de poder de processamento não implica em aumento relevante de performance, o que condiz com o esperado dadas as características da aplicação [40]. Com isso, reforça-se a hipótese de que as máquinas preemptáveis de categorias mais econômicas entregam melhor custo-benefício na execução do Sysbench *fileio*.

4.2.2 Simulação em Larga Escala

Constatado o baixo impacto da variação de vCPUs e de memória na performance do Sysbench *fileio*, foi investigado também o impacto de uma variação na dimensão do espaço

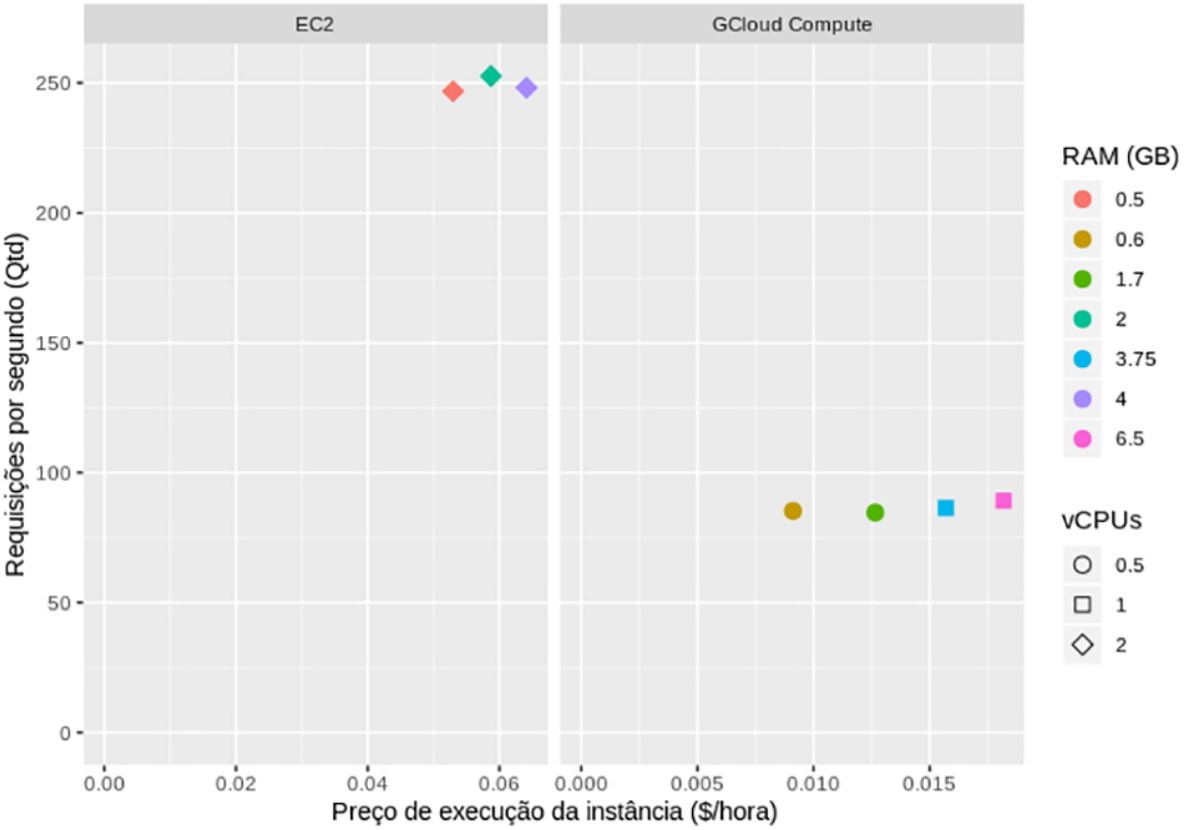

Figura 4.15: Custo e Desempenho Sysbench *fileio* por provedor.

de trabalho percorrido pelas leituras e escritas do *benchmark*. A análise em questão foi feita com base na simulação de uma carga de trabalho composta por um bilhão de operações do Sysbench *fileio*. Essas métricas são similares aos testes realizados com o *benchmark* SPECjvm2008, mas o número de operações realizadas foi aumentado para melhor visualização. A Figura 4.16 apresenta os resultados obtidos nos novos testes em ambos os provedores. O *eixo X* dos gráficos apresenta o tempo de execução da carga, enquanto o *eixo Y* descreve o custo em dólar para a execução da mesma. As amostras anteriormente apresentadas, cujo domínio de aplicação é de 80 GB (Figura 4.15), são dispostas para comparação com máquinas de mesmo hardware, mas com domínio dez vezes menor, de apenas 8 GB (Figura 4.16).

Os gráficos da Figura 4.16 apresentam uma performance superior em todos os registros cujo espaço de trabalho foi reduzido, com ganho de desempenho ainda maior nas instâncias da Google. Além disso, é possível constatar que a variação no tempo de execução foi maior nas instâncias com mais RAM disponível, o que contraria a leitura feita anteriormente de que a provisão de memória adicional não traz benefícios para a performance. Segundo a Wiki do Sysbench [40], esse ganho se dá devido a processos de *file caching* nos quais

operações I/O são feitas em memória, e não diretamente no dispositivo de armazenamento. Assim, conclui-se que o impacto da memória no desempenho do Sysbench *fileio* depende de uma proporção adequada entre a quantidade de RAM disponível e o tamanho do espaço de trabalho percorrido pelas leituras e escritas da aplicação.

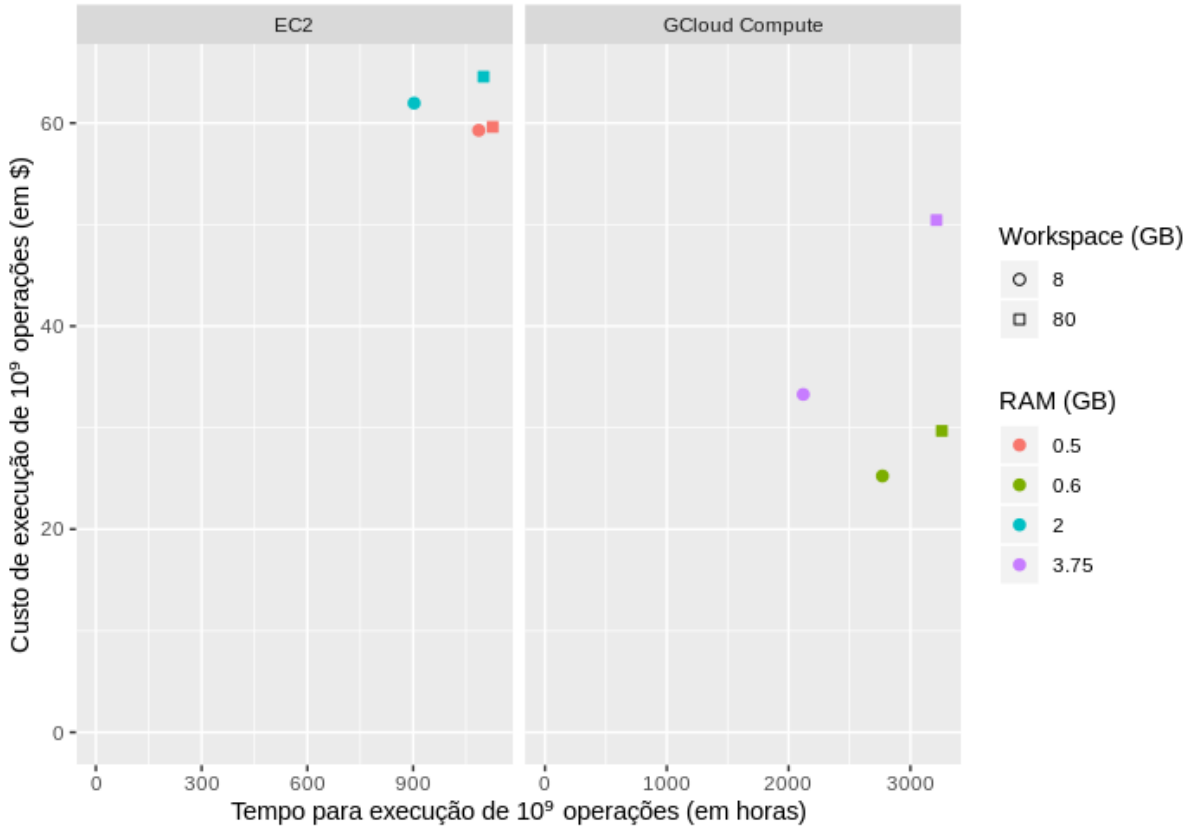

Figura 4.16: Domínio de aplicação Sysbench *fileio* e impacto em performance.

Diante do exposto e tomando uma perspectiva comparativa entre os provedores, foi realizada a simulação da carga de trabalho de um bilhão de operações para as máquinas apresentadas na Tabela 4.7, cujos resultados estão dispostos na Figura 4.17. Em leitura do gráfico (Figura 4.17), é perceptível que as instâncias *t3nano* e *small06* se mostram as mais indicadas para a execução do *fileio* dentre as instâncias preemptáveis de seus provedores, uma vez que as demais categorias apresentam custos mais elevados, com pouco ou nenhum ganho de performance.

Assim, colocando as instâncias dos dois provedores em comparação, fica evidente a dissonância entre o comportamento das máquinas em questão. Ainda que a máquina *t3nano* (Amazon) execute a carga de trabalho em aproximadamente um terço do tempo total obtido em *small06* (Google), seu custo é duplicado em relação à concorrente. Observa-se, dessa forma, um *tradeoff* entre desempenho e preço.

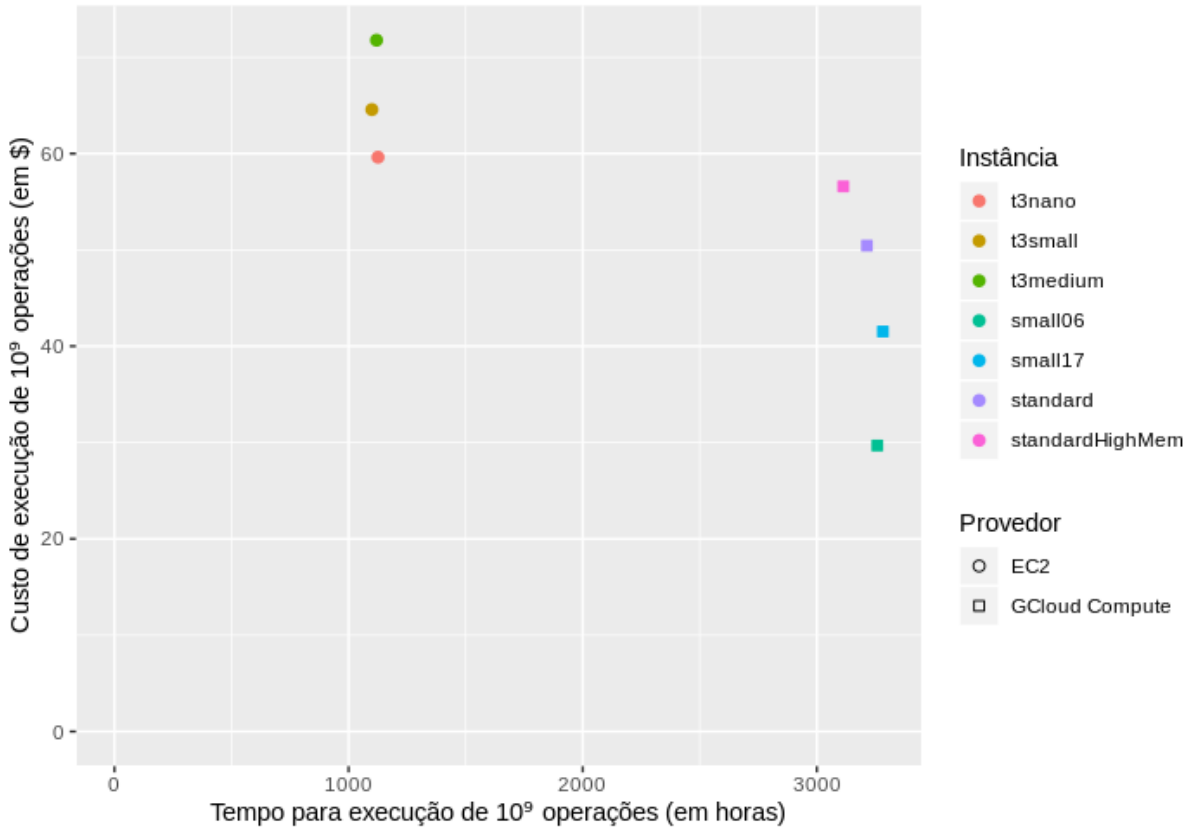

Figura 4.17: Execução de uma carga de trabalho Sysbench *fileio* com um bilhão de operações.

Dito isso, existem algumas características que podem explicar o comportamento revelado nos testes. A primeira delas é a tecnologia *EBS-Optimized Instances* da Amazon, na qual as máquinas de algumas famílias (T3, C5, entre outras) são providas com largura de banda dedicada entre suas instâncias e os discos provisionados no serviço EBS [29]. Essa funcionalidade torna a leitura e a escrita de dados em disco mais eficiente nessas instâncias, o que explica a performance superior obtida pelas instâncias Amazon no Sysbench *fileio*, visto que tal operação consiste no principal gargalo desse *benchmark*.

Identificada a causa para a performance superior, uma justificativa para o preço comparativamente maior nas instâncias Amazon é o custo atribuído às operações de leitura e de escrita realizadas no domínio EBS em discos magnéticos (HDD) de categoria padrão, apresentado na Tabela 4.7. Ainda que o provedor ofereça outras categorias de disco HDD que não contam com este custo adicional, elas não foram contempladas por este trabalho uma vez que não podem ser definidas como discos de inicialização *(root)* da instância. Os custos para execução das amostras da Figura 4.17 são apresentados na Tabela 4.8, de maneira que os valores atribuídos a máquinas virtuais preemptáveis, manutenção de disco

Tabela 4.8: Custos para execução da carga de trabalho Sysbench *fileio* simulada.

Provedor	**Instância**	**VMPrice (US$/hora)**	**IOPrice (US$/hora)**	**DiskPrice (US$/hora)**	**Total (US$/hora)**
Google	small06	0.003556 (39%)	0 (0%)	0.005556 (61%)	0.009111111 (100%)
Google	small17	0.007097 (56%)	0 (0%)	0.005556 (44%)	0.01265278 (100%)
Google	standard	0.010139 (65%)	0 (0%)	0.005556 (35%)	0.01569444 (100%)
Google	standard HighMem	0.012625 (69%)	0 (0%)	0.005556 (31%)	0.01818056 (100%)
Amazon	t3nano	0.0016 (3%)	0.0444186 (84%)	0.00694 (13%)	0.05296304 (100%)
Amazon	t3small	0.0063 (11%)	0.0454734 (77%)	0.00694 (12%)	0.05871784 (100%)
Amazon	t3medium	0.0125 (19%)	0.044667 (70%)	0.00694 (11%)	0.06411144 (100%)

e operações de leitura e de escrita estão discretizados.

Tendo essas características em mente, é possível constatar que nenhum dos provedores possui dominância absoluta de custo-benefício para a execução do Sysbench *fileio*, ou seja, cada provedor é superior em relação à uma métrica diferente.

4.2.3 Dispositivos SSD e seus Impactos

Uma outra configuração personalizável durante a criação de instâncias preemptáveis é o dispositivo utilizado para armazenamento, seja ele um disco rígido (HDD) ou uma unidade de estado sólido (SSD). Segundo a documentação do EBS [29], os SSDs são os equipamentos mais recomendados para cargas de trabalho com leitura e escrita de dados em grandes volumes ou intensiva, como é o caso do Sysbench *fileio*. Essa recomendação parte da característica de tais equipamentos utilizarem memória *flash* em vez de discos magnéticos - apesar do custo por unidade de memória ser mais elevado nestes dispositivos, o tempo de resposta de memórias *flash* é drasticamente menor [54]. A Figura 4.18 conta com os resultados mais interessantes já obtidos por máquinas HDD adicionados ao desempenho registrado por instâncias similares equipadas com discos SSD.

Em relação à oferta de disco, a Amazon conta com duas configurações de discos SSD: uma de uso intensivo *(IO1)* e outra de uso genérico *(GP2)*. Uma característica interessante dos discos *IO1* é a possibilidade de provisionar performance de recursos para garantir performance de operações I/O, sendo esta última configurável com até 50 IOPS por GB do disco em questão. O disco *GP2*, em contrapartida, não possui custos atrelados às suas

operações de leitura e escrita, os quais foram de grande impacto na comparação realizada entre os discos HDD dos provedores (Figura 4.17).

Em leitura do gráfico (Figura 4.18), fica evidente o ganho econômico e performático obtido com os discos SSD. Os resultados obtidos pelas execuções *SSD-GP2* da Amazon registram queda de custo em mais de 60% e queda de tempo total de execução em pelo menos 50%, quando comparados com os resultados de execuções HDD. Para a Google, esses números superam os 70%.

Sob a perspectiva de tempo de resposta, a Figura 4.18 sugere que as instâncias Google com unidades SSD e Amazon com dispositivos *SSD-IO1* possuem os melhores resultados, com tempo de execução consideravelmente menor. Levando em consideração uma perspectiva financeira, entretanto, a simulação realizada para o disco *SSD-IO1* trouxe resultados pouco interessantes, uma vez que esta configuração atribui um custo elevado à provisão de operações I/O [29]. Com isso, conclui-se que as máquinas preemptáveis do provedor Google equipadas com discos SSD são as mais recomendadas para execução de cargas de trabalho com fluxo intenso de leitura e escrita em dispositivos, uma vez que apresentaram o menor custo e tempo de resposta dentre as execuções simuladas no *benchmark* Sysbench *fileio*.

Apresentados os resultados obtidos sob perspectiva do Sysbench *fileio*, a seção seguinte apresenta características dos provedores que não possuem impacto direto nos resultados quantitativos obtidos a partir das aplicações, mas são relevantes para possíveis tomadas de decisão referente à adoção de um destes provedores.

4.3 Considerações finais

Além das métricas de custo e tempo de performance, amplamente exploradas neste trabalho, existem outros aspectos relevantes para a escolha do provedor e instância preemptável mais adequados para um certo contexto. Algumas dessas características são de difícil quantificação e são abordadas nesta subseção.

Um dos aspectos importantes ainda não mencionados é a variação de preços ao qual as instâncias preemptáveis da Amazon estão sujeitas. Uma vez que sua precificação se dá por um modelo de leilão, podem ocorrer situações onde as despesas dos usuários de instâncias Spot sejam incompatíveis com expectativas prévias. A Amazon disponibiliza em sua plataforma um histórico com os preços das instância em cada região nos últimos três meses, o que pode auxiliar o usuário na tomada de decisão quanto ao seu lance máximo. Além disso, esse histórico facilita a comparação de preço entre instâncias Spot em relação às instâncias sob demanda, de uso dedicado. A título de exemplo, a Figura 4.19 mostra os preços registrados para a instância *c52xlarge* nas datas entre 28 de março e 24

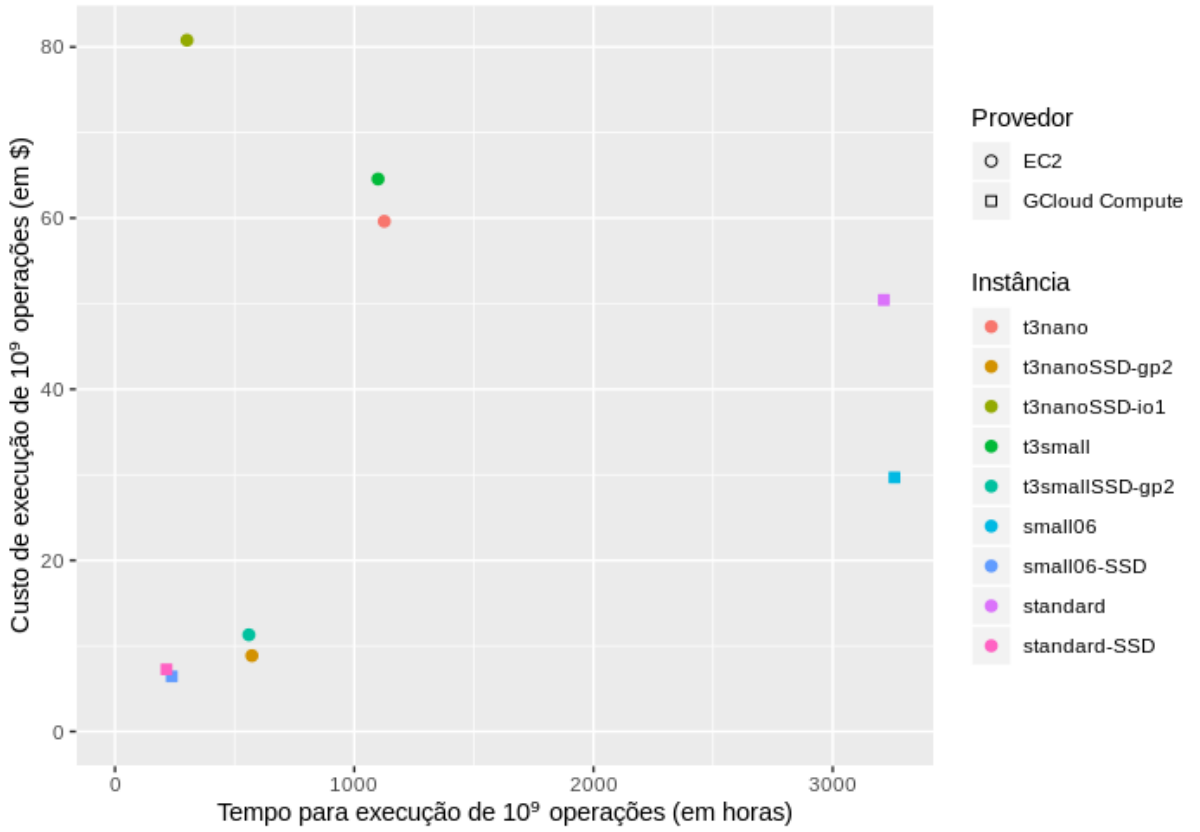

Figura 4.18: Execução de carga de trabalho Sysbench *fileio* incluindo instâncias equipadas com SSD.

de junho, período no qual os testes deste trabalho foram realizados. A leitura do gráfico, com baixa variação de preços no período em que os *benchmarks* foram executados, sugere que os valores de mercado relativamente estáveis do provedor na região recomendada (*us-east-2b*).

Outro fator que tem ganhado relevância com a expansão dos serviços de tipo IaaS para diversos usuários é a usabilidade, característica essa impactante não só para plataformas de nuvem, mas para qualquer aplicação ou sistema interativo. Usabilidade diz respeito à facilidade com que um usuário se familiariza com um serviço ou produto e consegue atingir seus objetivos com ele [55]. Diante disso, uma importante barreira de complexidade é o modelo de precificação Amazon apresentado neste trabalho, de maneira que o sistema baseado em leilões a cobrança sobre operações I/O dificulta a previsão de custos para usuários leigos. O modelo de cobrança do Google, em contrapartida, se mostra mais simples e acessível ao usuário comum uma vez que é baseado em valores fixos e pré-determinados antes do serviço ser utilizado.

Além do modelo de precificação, outro traço de usabilidade das plataformas é a maneira como são estabelecidas conexões entre uma máquina local e as instâncias em nuvem.

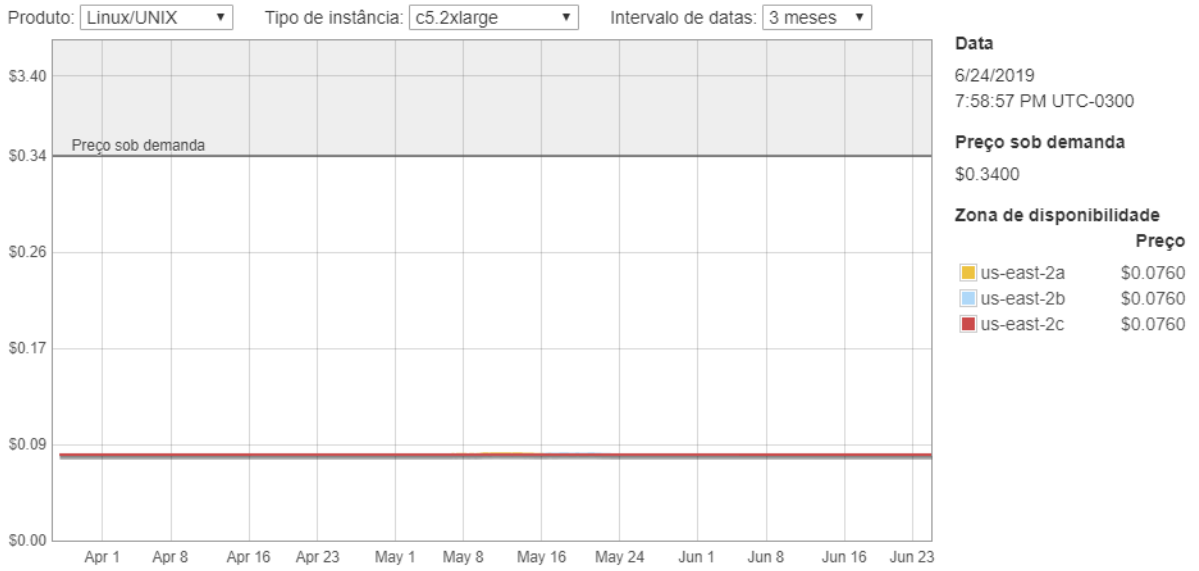

Figura 4.19: Histórico de preço de instâncias Spot *c52xlarge* na região *us-east-2b*.

Enquanto o provedor Amazon exige que o armazenamento de chaves privadas seja feito localmente, o Google oferece a possibilidade de adicionar essas informações aos metadados do projeto, o qual é vinculado à conta Google do usuário. Dessa maneira, o provedor disponibiliza ao usuário uma maneira mais simples e flexível de acessar as máquinas virtuais Google, através de uma plataforma web independente de *plugins* (Figura 4.20) [56].

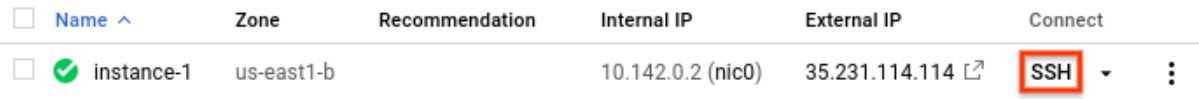

Figura 4.20: Acesso facilitado a instâncias Google.

Diante disso, conclui-se que a plataforma de nuvem Google apresenta uma série de vantagens para usuários com pouca ou nenhuma experiência na gestão de máquinas virtuais, característica essa que pode ter um grande peso na escolha de provedores por parte de consumidores de nuvem. Apresentados os resultados, o capítulo seguinte descreve as principais conclusões atingidas com este trabalho e sugestões para trabalhos futuros.

Capítulo 5

Conclusões

Este trabalho propôs uma avaliação dos serviços de máquinas preemptáveis oferecidos pelos provedores de nuvem pública Amazon e Google. O foco da avaliação foi analisar custo e performance a partir do interesse do usuário de usufruir de um serviço adequado pelo menor preço. Para tal, foram desenvolvidos testes experimentais com *benchmarks* executados em diversas instâncias preemptáveis desses provedores, produzindo dados para uma análise quantitativa.

Em relação à perspectiva dos resultados nos testes com os *benchmarks* SPECjvm2008, foi constatado que a performance de uma instância preemptável está diretamente relacionada ao número de vCPUs. Com isso, as máquinas de núcleo compartilhado oferecidas pelo provedor Google registraram um custo e tempo de execução muito inferiores às demais. Quanto a personalização de processadores oferecida pela Amazon, as máquinas provisionadas com hardware de ambos os fabricantes disponibilizados, Intel e AMD, apresentaram superioridade em *benchmarks* específicos.

Para trabalhos futuros, sugere-se que sejam investigadas características das cargas de trabalho e dos processadores que expliquem a performance superior de um fabricante em relação ao outro.

Ainda para o SPECjvm2008, o custo mínimo para a execução da carga de trabalho foi obtido por máquinas virtuais de pequeno porte, sendo elas das categorias *t3small* e standardHighCPU nos provedores Amazon e Google, respectivamente. Sob uma perspectiva que priorize tempo de resposta, as instâncias das categorias *HighCPU* do Google e *C5* da Amazon registraram melhor custo-benefício, com tempo de execução inversamente proporcional ao número de vCPUs.

Assim sendo, em comparação direta entre os provedores, as instâncias preemptáveis instanciadas pela Amazon apresentaram resultados com performance superior, e custo reduzido em diversas faixas de preço, ou seja, as máquinas preemptáveis da Amazon

se mostraram mais indicadas para execução de cargas de trabalho com processamento intensivo, como o SPECjvm2008.

Por outro lado, analisando os resultados do *benchmark* Sysbench *fileio* em máquinas preemptáveis, observou-se que a adição de vCPUs e memória RAM impactaram pouco na performance da execução. Com isso, as categorias de máquinas mais indicadas para execução intensiva de operações I/O aleatórias são as mais econômicas, sendo essas as famílias *t3nano* na Amazon e *small06* no Google.

No domínio de máquinas com discos HDD, a Amazon apresentou uma performance consideravelmente superior na execução do Sysbench *fileio*, mas acompanhada de um aumento proporcional nos custos.

Com isso, nenhum dos provedores registrou superioridade absoluta. Considerando a execução do *benchmark* em instâncias equipadas com dispositivos SSD, em contrapartida, as instâncias Google se mostraram superiores em relação às máquinas Amazon, tanto em relação ao tempo de execução quanto financeiramente. Com isso, conclui-se que as instâncias Google equipadas com SSD são as mais recomendadas para cargas de trabalho limitadas pela leitura e escrita aleatória em dispositivos de armazenamento.

Em relação à usabilidade, concluiu-de que o provedor Google apresenta características e interfaces que facilitam o uso e o planejamento de custos das máquinas preemptáveis. Essa constatação ganha relevância no que se refere a usuários inexperientes, com pouco ou nenhum conhecimento acerca de protocolos de segurança e gerenciamento remoto.

Além disso, as preempções diárias previstas pelas máquinas preemptáveis do Google podem trazer impactos negativos ao cliente do serviço, enquanto as instâncias *Spot* da Amazon têm sua execução garantida com maior grau de confiabilidade, caso o usuário esteja disposto a pagar mais caro quando necessário. Essas características, vinculadas aos modelos de precificação praticados pelas companhias, não podem ser ignoradas na tomada de decisão sobre qual dos serviços adotar. Por fim, conclui-se que o conhecimento da carga de trabalho também é de suma importância para que o usuário selecione o provedor e o tipo de instância que entreguem o melhor custo-benefício segundo suas necessidades.

Para trabalhos futuros, sugere-se que sejam executados testes considerando outros provedores, como Microsoft Azure e Oracle Cloud. Além disso, é interessante avaliar o uso de GPUs em máquinas preemptáveis e seus consequentes impactos. Propõe-se também a realização de experimentos com outros *benchmarks*, cujas cargas de trabalho tenham caracteristicas diferentes do SPECjvm2008 e do Sysbench *fileio*. Alguns exemplos sugeridos são testes com aplicações diretamente impactadas pela memória RAM disponível, ou pela latência de comunicação entre um usuário cliente e um servidor em nuvem, onde a região geográfica da instância ganha maior relevância.

Por fim, também sugere-se a análise comparativa entre máquinas preemptáveis e máquinas dedicadas, buscando maneiras de estimar o custo e o benefício vinculados à confiabilidade oferecida pelas máquinas sob demanda.

Referências

[1] *Gartner forecasts worldwide public cloud revenue to grow 17.3 percent in 2019.* https://www.gartner.com/en/newsroom/press-releases/. Acessado em: 2019-06-25. 1

[2] *Gartner magic quadrant for cloud infrastructure as a service 2018.* https://www.bmc.com/blogs/gartner-magic-quadrant-cloud-iaas/. Acessado em: 2019-06-25. 2

[3] *Announcing low-priority vms on scale sets now in public preview.* https://azure.microsoft.com/pt-br/blog/low-priority-scale-sets/. Acessado em: 2019-07-22. 2

[4] Giordanelli, Raffaele e Carlo Mastroianni: *The cloud computing paradigm: Characteristics, opportunities and research issues.* Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR), 2010. 5, 11

[5] Buyya, Rajkumar, Chee Shin Yeo, Srikumar Venugopal, James Broberg e Ivona Brandic: *Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility.* Future Generation computer systems, 25(6):599–616, 2009. 5

[6] Mell, Peter, Tim Grance *et al.*: *The nist definition of cloud computing.* 2011. 5, 7

[7] Lakew, Ewnetu Bayuh, Cristian Klein, Francisco Hernandez-Rodriguez e Erik Elmroth: *Towards faster response time models for vertical elasticity.* Em *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, páginas 560–565. IEEE, 2014. 6

[8] Hashem, Ibrahim Abaker Targio, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani e Samee Ullah Khan: *The rise of "big data" on cloud computing: Review and open research issues.* Information systems, 47:98–115, 2015. 6

[9] Jadeja, Yashpalsinh e Kirit Modi: *Cloud computing-concepts, architecture and challenges.* Em *2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET)*, páginas 877–880. IEEE, 2012. 6

[10] *Rightscale 2019 state of the cloud report.* https://info.flexerasoftware.com/SLO-WP-State-of-the-Cloud-2019. Acessado em: 2019-05-26. 8, 9

[11] Al-Roomi, May, Shaikha Al-Ebrahim, Sabika Buqrais e Imtiaz Ahmad: *Cloud computing pricing models: a survey*. International Journal of Grid and Distributed Computing, 6(5):93–106, 2013. 8, 9

[12] Weinhardt, Christof, Arun Anandasivam, Benjamin Blau, Nikolay Borissov, Thomas Meinl, Wibke Michalk e Jochen Stößer: *Cloud computing – a classification, business models, and research directions*. Business & Information Systems Engineering, 1(5):391–399, Oct 2009, ISSN 1867-0202. `https://doi.org/10.1007/s12599-009-0071-2`. 8

[13] Baset, Salman A: *Cloud slas: present and future*. ACM SIGOPS Operating Systems Review, 46(2):57–66, 2012. 9

[14] *Amazon compute service level agreement*. `https://aws.amazon.com/compute/sla/`. Acessado em: 2019-05-26. 9

[15] *Google compute engine service level agreement*. `https://cloud.google.com/compute/sla`. Acessado em: 2019-05-26. 9

[16] Zhang, Qi, Quanyan Zhu e Raouf Boutaba: *Dynamic resource allocation for spot markets in cloud computing environments*. Em *2011 Fourth IEEE International Conference on Utility and Cloud Computing*, páginas 178–185. IEEE, 2011. 10, 14

[17] *Aws spot instances documentation*. `https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-spot-instances.html`. Acessado em: 2019-06-06. 10

[18] *Google preemptible vm instances documentation*. `https://cloud.google.com/compute/docs/instances/preemptible`. Acessado em: 2019-06-06. 10, 17, 24

[19] Subramanya, Supreeth, Tian Guo, Prateek Sharma, David Irwin e Prashant Shenoy: *Spoton: a batch computing service for the spot market*. Em *Proceedings of the sixth ACM symposium on cloud computing*, páginas 329–341. ACM, 2015. 10, 14

[20] Ekanayake, Jaliya e Geoffrey Fox: *High performance parallel computing with clouds and cloud technologies*. Em *International Conference on Cloud Computing*, páginas 20–38. Springer, 2009. 11

[21] Appel, Stefan, Ilia Petrov e Alejandro Buchmann: *Performance evaluation of multi machine virtual environments*. Em *2010 SPEC Benchmark Workshop*, páginas 1–13, 2010. 11, 20

[22] Li, Ang, Xiaowei Yang, Srikanth Kandula e Ming Zhang: *Comparing public-cloud providers*. IEEE Internet Computing, 15(2):50–53, 2011. 12, 13

[23] *Specjvm2008 user guide*. `https://www.spec.org/jvm2008/docs/UserGuide.html`. Acessado em: 2019-05-26. 13, 18

[24] *Sysbench manual*. `http://imysql.com/wp-content/uploads/2014/10/sysbench-manual.pdf`. Acessado em: 2019-06-21. 13

[25] *Big four still dominate in q1 as cloud market growth exceeds 50%*. https://www.srgresearch.com/articles/gang-four-still-racing-away-cloud-market. Acessado em: 2019-05-26. 13

[26] *The cloud wars explained: Amazon is dominating, but microsoft and google are striking back*. https://finance.yahoo.com/news/cloud-wars-explained-amazon-dominating-120000974.html. Acessado em: 2019-05-26. 13

[27] *Amazon company*. https://en.wikipedia.org/wiki/Amazon_(company). Acessado em: 2019-06-24. 13

[28] *Timeline of amazon web services*. https://en.wikipedia.org/wiki/Timeline_of_Amazon_Web_Services. Acessado em: 2019-06-15. 13, 15

[29] *Amazon ebs features*. https://aws.amazon.com/ebs/features/. Acessado em: 2019-06-13. 13, 42, 43, 44

[30] *Timeline of amazon web services*. https://aws.amazon.com/pt/ec2/instance-types/?nc1=h_ls. Acessado em: 2019-06-15. 13

[31] Taifi, Moussa, Justin Y Shi e Abdallah Khreishah: *Spotmpi: a framework for auction-based hpc computing using amazon spot instances*. Em *International Conference on Algorithms and Architectures for Parallel Processing*, páginas 109–120. Springer, 2011. 14

[32] Yi, Sangho, Derrick Kondo e Artur Andrzejak: *Reducing costs of spot instances via checkpointing in the amazon elastic compute cloud*. Em *2010 IEEE 3rd International Conference on Cloud Computing*, páginas 236–243. IEEE, 2010. 14, 15, 21

[33] *Google app engine*. https://cloud.google.com/appengine/. Acessado em: 2019-06-24. 15

[34] *Google compute engine*. https://cloud.google.com/compute. Acessado em: 2019-06-24. 15

[35] *Google compute engine*. https://en.wikipedia.org/wiki/Google_Compute_Engine. Acessado em: 2019-06-15. 15

[36] *Tipos de máquinas google*. https://cloud.google.com/compute/docs/machine-types?hl=pt-br#sharedcore. Acessado em: 2019-06-15. 16

[37] *Benchmark definition*. https://en.wikipedia.org/wiki/Benchmark_(computing). Acessado em: 2019-06-24. 17

[38] Oi, Hitoshi: *A comparative study of jvm implementations with specjvm2008*. Em *2010 Second International Conference on Computer Engineering and Applications*, volume 1, páginas 351–357. IEEE, 2010. 18, 21

[39] *Oracle java archive*. https://www.oracle.com/technetwork/java/archive-139210.html. Acessado em: 2019-06-20. 19

[40] *Sysbench - gentoo wiki*. https://wiki.gentoo.org/wiki/Sysbench. Acessado em: 2019-06-13. 19, 39, 40

[41] Garg, Saurabh Kumar, Steve Versteeg e Rajkumar Buyya: *Smicloud: A framework for comparing and ranking cloud services*. Em *2011 Fourth IEEE International Conference on Utility and Cloud Computing*, páginas 210–218. IEEE, 2011. 20

[42] Lengauer, Philipp, Verena Bitto, Hanspeter Mössenböck e Markus Weninger: *A comprehensive java benchmark study on memory and garbage collection behavior of dacapo, dacapo scala, and specjvm2008*. Em *Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering*, páginas 3–14. ACM, 2017. 20

[43] Juve, Gideon, Ewa Deelman, G Bruce Berriman, Benjamin P Berman e Philip Maechling: *An evaluation of the cost and performance of scientific workflows on amazon ec2*. Journal of Grid Computing, 10(1):5–21, 2012. 20

[44] Tatlow, PJ e Stephen R Piccolo: *A cloud-based workflow to quantify transcript-expression levels in public cancer compendia*. Scientific reports, 6:39259, 2016. 21

[45] *Specjvm2008 frequently asked questions*. https://www.spec.org/jvm2008/docs/FAQ.html. Acessado em: 2019-06-15. 24

[46] *Specjvm2008 benchmarks*. https://www.spec.org/jvm2008/docs/benchmarks/index.html. Acessado em: 2019-06-15. 26

[47] *Bottleneck (software)*. https://en.wikipedia.org/wiki/Bottleneck_(software). Acessado em: 2019-06-13. 29

[48] *Aws ec2 instance comparison: C4 vs c5*. https://www.learnaws.org/2017/11/17/comparing-ec2-c4-c5/. Acessado em: 2019-06-13. 31

[49] *Amazon ec2 regions and availability zones*. https://docs.aws.amazon.com/pt_br/AWSEC2/latest/UserGuide/using-regions-availability-zones.html. Acessado em: 2019-06-13. 32, 33

[50] *Do ec2 prices depend on the region?* https://www.awsforbusiness.com/ec2-prices-depend-region/. Acessado em: 2019-06-13. 32

[51] *Google cloud regions and zones*. https://cloud.google.com/compute/docs/regions-zones/. Acessado em: 2019-06-13. 33

[52] *Save yourself a lot of pain (and money) by choosing your aws region wisely*. https://www.concurrencylabs.com/blog/choose-your-aws-region-wisely/. Acessado em: 2019-06-13. 33

[53] *Coeficiente de variação - wikipedia*. https://en.wikipedia.org/wiki/Coefficient_of_variation. Acessado em: 2019-06-15. 36

[54] *Solid-state drive*. https://en.wikipedia.org/wiki/Solid-state_drive. Acessado em: 2019-07-22. 43

[55] *Usability definition*. https://www.interaction-design.org/literature/topics/usability. Acessado em: 2019-06-21. 45

[56] *Como estabelecer conexão com instâncias*. https://cloud.google.com/compute/docs/instances/connecting-to-instance?hl=pt-br. Acessado em: 2019-06-21. 46