Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 27, Num. 04 (2020) 45-62

RESEARCH ARTICLE

Cloud infrastructure for HPC investment analysis

Análise de investimento em infraestrutura de nuvem para HPC

Maicon Ança dos Santos[1]*, Gerson Geraldo H. Cavalheiro[1]

Resumo: With the consolidation of cloud computing technology, there is a growing interest in exploring it to support High Performance Computing (HPC). However, migrating such applications to public or private cloud environments brings some challenges, in particular, the cost in financing the migration process. In this paper, a literature review is presented with selected papers about analyzing cloud infrastructure investments. In particular, the selected papers analyse how investments impact applications. For discussion of related works, conditions for running HPC applications in the cloud are characterized.
Keywords: Cloud computing — HPC — Cost model — Investment

Resumo: Com a consolidação da tecnologia de computação em nuvem, cresce o interesse em explorá-las para oferecer suporte à computação de alto desempenho (HPC). No entanto, migrar essas aplicações para ambientes de nuvem, públicos ou privados, traz alguns desafios, em particular, o custo que envolve o financiamento do processo de migração. Neste artigo, é apresentada uma revisão sistemática da literatura na qual foram identificados trabalhos relacionados à análise de investimentos em infraestruturas de nuvem. Em particular, os trabalhos de interesse são aqueles que analisam o impacto nas aplicações HPC a partir dos investimentos realizados. Para discussão de trabalhos relacionados, são caracterizadas condições para execução de aplicações HPC em nuvem.
Palavras-Chave: Computação em Nuvem — HPC — Modelo de Custo — Investimento

[1] *Universidade Federal de Pelotas, Pelotas, Rio Grande do Sul, Brasil*
*__Corresponding author__: madsantos@inf.ufpel.edu.br
DOI: http://dx.doi.org/10.22456/2175-2745.106794 • **Received:** 24/08/2020 • **Accepted:** 23/10/2020

1. Introdução

A computação em nuvem é uma realidade tanto em ambientes acadêmicos quanto comerciais. Este modelo de computação se apresenta como uma infraestrutura de suporte à execução sob demanda para aplicações que vão desde simples portais web a grandes fluxos de trabalho científicos. Neste ambiente computacional, recursos podem ser rapidamente provisionados e liberados com o mínimo esforço de gerenciamento ou interação com o provedor de serviços [1]. Caso a decisão para implantar uma solução implique no uso de nuvens computacionais, soluções utilizando nuvens públicas ou privadas devem ser consideradas. Neste trabalho é realizado uma Revisão Sistemática da Literatura (RSL) em que é tratado o tema referente ao dimensionamento de custos da adoção de nuvens computacionais no contexto do processamento de alto desempenho.

O provisionamento de recursos computacionais em uma nuvem é realizado de forma elástica, em função da demanda da aplicação. O modelo de implementação de nuvem pública oferece elasticidade por meio de um modelo de pagamento por utilização (*pay-as-you-go*), no qual a alocação de recursos aumenta ou diminui à medida que o usuário final disponibiliza mais ou menos recursos financeiros. Neste modelo, as demandas de processamento e disponibilidade, que antes faziam parte de estruturas locais, são repassadas a provedores externos responsáveis por garantir os acessos necessários, desonerando o usuário do laborioso processo de gestão dos recursos.

Por outro lado, nas implementações de nuvens privadas, os recursos computacionais aumentam ou diminuem, conforme a demanda, com a ativação ou desativação de nós de processamento, sem a necessidade de aportes financeiros por parte dos usuários durante as operações. Nestes ambientes também é possível programar o uso dos recursos em função da demanda. Com a diminuição dos custos de implementação das estruturas de nuvem, a opção por implantar *data centers* próprios como suporte às nuvens privadas é recorrente em instituições, comerciais ou acadêmicas. Empresas e cooperativas, como a Unimed[1], já optaram por migrar seus serviços, antes alocados em infraestruturas de terceiros, para dentro de seus ambientes locais. No contexto acadêmico, como exemplo, a Universidade de São Paulo (USP) disponibiliza, tanto

[1] <http://www.unimed.coop.br/portalunimed/relatorio2013/realizacoes.html>. Acesso em: 15 outubro 2020.

para comunidade acadêmica quanto comunidade externa, o projeto InterNuvem[2], o qual oferece acesso, por parte de pesquisadores, a serviços de armazenamento e processamento de dados de alto desempenho em nuvens computacionais interconectadas. Já a empresa Tecnologia Bancária (TecBan)[3], responsável pela rede de caixas eletrônicos Banco24Horas, utiliza infraestrutura de nuvem privada para hospedar sistemas críticos, que necessitam de maior garantia quanto à confiabilidade e segurança, exigidos pelo modelo de negócio ao qual pertence.

Características como rápido provisionamento de recursos, escalabilidade, recuperação de desastres, manutenções centralizadas de hardware e software fazem das nuvens ambientes favoráveis à migração de aplicações locais, executadas em *clusters* dedicados, para estruturas hospedadas em provedores de nuvens públicas ou em nuvens privadas. Porém, é importante ter clareza quanto aos recursos necessários por cada aplicação a fim de que se possa extrair os melhores benefícios financeiros de cada proposta.

Com a popularização e constante evolução das nuvens computacionais, demandas pela execução de aplicações que exigem alto poder de processamento se tornam mais presentes, oportunizando frentes de pesquisa sobre a viabilidade de execuções *High Performance Computing* (HPC) em nuvens (públicas ou privadas). Consequentemente, também são realizados esforços objetivando identificar os tipos de aplicações que terão melhor desempenho em ambientes altamente distribuídos e as consequências financeiras na adoção de soluções. Para os provedores de serviço de nuvem, o principal objetivo é a entrega de conteúdo e não aplicações que demandem alto poder de processamento. Com isso, é necessário analisar o comportamento das aplicações e o cenário de uso pretendido, a fim de que seja possível a escolha da infraestrutura mais adequada para execução [2].

As aplicações HPC possuem características específicas que expressam suas necessidades computacionais. Além de necessidades típicas ligadas à capacidade do processador e da disponibilidade de memória, em sistemas distribuídos requisitos ligados à comunicação também são determinantes para verificar sua adequação aos ambientes de nuvem. O uso de elementos de hardware não apropriados para alto desempenho e sobrecargas geradas pelos processos de virtualização constituem obstáculos na adoção de nuvens por parte dos usuários de computação de alto desempenho (HPC). Por exemplo, aritméticas de ponto flutuante, comunicação entre processos e até mesmo a escolha de *drivers* para máquinas virtuais podem afetar, significativamente, o desempenho do processo computacional [3]. Com relação às decisões de escalonamento, estas podem afetar o desempenho das tarefas a serem executadas. Escalonadores compatíveis com HPC podem melhorar o desempenho das aplicações em ambientes de nuvem, pois conseguem explorar as propriedades da infraestrutura e das próprias aplicações [4]. Assim, algumas questões relacionadas a quem se beneficiaria com a migração para ambientes de nuvem, qual, por que e como cada aplicação HPC pode ser submetida, devem ser consideradas.

Neste artigo, o enfoque está na análise do investimento em uma infraestrutura de nuvem para aplicações com alta demanda de processamento. A revisão sistemática da literatura, portanto, coleta trabalhos relacionados à determinação de modelos de custos financeiros associados a implantação e gestão de infraestruturas de nuvem. Em uma situação ideal, o usuário obtém o desempenho desejado realizando o mínimo de investimento em infraestrutura. Assim, não existe nem sub nem superdimensionamento da infraestrutura e as consequentes perdas financeiras decorrentes, respectivamente, do baixo índice de produção ou do passivo construído em hardware não explorado. O objetivo, nesta RSL, é identificar abordagens que relacionem perdas e ganhos financeiros da solução adotada quanto ao desempenho obtido no suporte à aplicação. Como resultado, são elencados os trabalhos mais relevantes no contexto de análise de custos de implantação de nuvens que suportem a execução de aplicações HPC. De posse deste material é possível analisar, sob diferentes perspectivas, se as ofertas de computação em nuvem (Infraestrutura como Serviço, principalmente) são capazes de atender os diferentes tipos de demandas, levando em consideração os custos envolvidos nas operações. Nesta abordagem são considerados tanto as soluções que adotam nuvens privadas quanto nuvens públicas.

As contribuições deste trabalho são: *a*) a própria seleção de trabalhos que relacionam análise do custo de adoção de infraestruturas de nuvem para suporte à aplicações HPC; *b*) a sumarização da abordagem de cada um dos trabalhos selecionados; e, por fim *c*) a identificação de oportunidades de pesquisa não tratadas nos trabalhos elencados, estando em aberto para investigação.

O texto deste artigo está organizado da seguinte forma: na Seção 2 é apresentado o contexto da pesquisa, caracterizando os requisitos de execução para aplicações HPC em ambientes de nuvens computacionais e as métricas de custos financeiros para tais execuções. A Seção 3 apresenta a revisão sistemática da literatura com o objetivo de identificar o estado da arte na execução de aplicações HPC em nuvem, juntamente com os custos financeiros envolvidos. Uma sumarização dos trabalhos relacionados ao tema é apresentado na Seção 4 e a síntese da leitura destes trabalhos, respondendo as Questões de Pesquisa elaboradas para a presente RSL, na Seção 5. Por fim, a Seção 6, encerra a leitura deste artigo, tecendo algumas considerações sobre o tema estudado.

2. Requisitos de Execução para Aplicações HPC

Em ambientes de nuvem, a precificação é o processo pelo qual se determina quanto um provedor de serviços receberá de um usuário final pelos serviços prestados [5]. O processo de

[2] <https://cetisp.sti.usp.br/competencias/internuvem/>. Acesso em: 15 outubro 2020.

[3] <http://epocanegocios.globo.com/Revista/Common/0,,EMI96465-17453,00-NUVEM+PUBLICA+OU+PRIVADA.html>. Acesso em: 15 outubro 2020.

precificação pode ser fixo, no qual o usuário final paga sempre o mesmo valor durante todo o tempo de uso dos serviços; dinâmico, em que os valores cobrados mudam com base em alterações de características; e dependente de mercado, que é quando o cliente é cobrado em função das condições de mercado em tempo real.

Nesta seção são apresentados diferentes custos associados à implantação de nuvens, Seção 2.1, assim como caracterizados os aspectos que influenciam nas suas respectivas variações, Seção 2.2. O custo da implantação de uma nuvem é contabilizado a partir da composição destes custos individuais, Seção 2.3. Este trabalho, tendo interesse na exploração de nuvens para HPC, apresenta uma abordagem particular para este caso específico na Seção 2.4. A discussão final, Seção 2.5, aborda o contexto de custos analisado sob a ótica das necessidades de implantação de nuvens para HPC para delimitar o contexto da Revisão Sistemática da Literatura conduzida nesta pesquisa.

2.1 Fontes de Custos em IaaS

De um modo geral, os recursos disponibilizados por nuvens do tipo IaaS são difíceis de ser quantificados, fazendo com que seja necessária a utilização de modelos de decisão orientados a custos. Um dos modelos mais utilizados é o custo total de propriedade – TCO (*Total Cost of Ownership*) [6].

O custo total de propriedade, inicialmente definido pela empresa Gartner, Inc.[4], é reconhecido como o método padrão para a análise financeira de recursos de Tecnologia da Informação (TI) e outros custos empresariais relacionados à TI [7]. O modelo de TCO inclui aquisição, gerenciamento e suporte de hardware e software, comunicações, despesas do usuário final e os custos com tempo de inatividade, treinamento e outras perdas de produtividade. Para Filiopoulou e seu grupo [8], a estimativa do custo total de propriedade, em particular, é um procedimento que fornece os meios para determinar o valor econômico total de um investimento, incluindo as despesas iniciais de capital (CAPEX, *capital expenditure*) e as despesas operacionais (OPEX, *operational expenditure*). No contexto de computação em nuvem, corresponde à estimativa de valores necessários para implementação e operação de uma infraestrutura de nuvem.

Os benefícios do uso da abordagem de TCO estão na melhoria da comunicação entre cliente e provedor e na análise de todo o ciclo de vida dos artefatos de TI. Além disso, é possível analisar os custos ou componentes de custos individuais de um artefato de TI por meio de um esquema predefinido [9]. Como o objetivo do modelo TCO é fornecer uma visão abstrata e simplificada do "mundo real", em vez de incluir todos os custos relevantes na análise, a complexidade da realidade pode ser reduzida trabalhando com base em premissas e incluindo apenas um número limitado de fatores de custo cuidadosamente selecionados.

Métricas para os cálculos de custo e investimento em infraestruturas de nuvens computacionais necessitam de um modelo claro, juntamente com os fatores que influenciam estes custos. A Tabela 1 atribui fatores de custo f para cada tipo de custo t identificado. Os tipos de custo $t \in T$ e os fatores de custo $f \in F$ estão sujeitos aos conjuntos T e F:

$$T = \{dEst, ava, txIaaS, imp, sup, trein, manut, falha, bs\}$$

$$F = \{dTempo, sCon, itDec, pComp, cArm, tEnt, tSda, tInt, nCon, dom, ssl, lic, txServ, pPort, cSup, rProb, tPrep, tPart, mInstr, perda\}$$

Cada fator influenciador é atribuído a um tipo de custo e, em seguida, os elementos dos conjuntos T e F são agrupados em fórmulas. Cada fórmula é aplicada a um custo específico, possibilitando a obtenção dos valores para cada tipo de custo.

Tabela 1. Tipos de custo e fatores de custo relacionados

Tipo de custo	Fatores de custo
Decisão estratégica, seleção de serviços de computação em nuvem e tipos de nuvem (*dEst*)	Despesas de tempo (*dTempo*), serviços de consultoria (*sCon*), informações para tomada de decisão (*itDec*).
Avaliação e seleção de prestador de serviços (*ava*)	Despesas de tempo (*dTempo*), serviços de consultoria (*sCon*), informações para tomada de decisão (*itDec*).
Taxa de serviço IaaS (*txIaaS*)	Poder de computação (*pComp*), capacidade de armazenamento (*cArm*), transferência de dados de entrada (*tEnt*), transferência de dados de saída (*tSda*), transferência de dados interna do provedor (*tInt*), número de consultas (*nCon*), domínio (*dom*), certificado SSL (*ssl*), licença (*lic*), taxa de serviço básica (*txServ*).
Implementação, configuração, integração e migração (*imp*)	Despesas de tempo (*dTempo*), processo de portabilidade (*pPort*).
Suporte (*sup*)	Despesas de tempo (*dTempo*), custos de suporte (*cSup*), resolução de problemas (*rProb*).
Treinamento inicial e permanente (*trein*)	Tempo de preparação dos funcionários internos (*tPrep*), tempo de participação dos funcionários internos (*tPart*), material de instruções (*mInstr*), serviços de consultoria externa (*sCon*).
Manutenção e modificação (*manut*)	Despesas de tempo (*dTempo*).
Falha no sistema (*falha*)	Perda por período (*perda*).
Backsourcing ou descarte (*bs*)	Despesas de tempo (*dTempo*), processo de portabilidade (*pPort*).

Fonte: Adaptado de [9].

2.2 Fatores de Custo

Os custos relacionados à decisão estratégica, seleção de serviços de computação em nuvem e tipos de nuvem (*dEst*), juntamente com os custos de avaliação e seleção de prestadores de serviço (*ava*) são influenciados pelas despesas de tempo (*dTempo*) necessário para a tomada de decisão, as despesas com informações nas quais a decisão pode ser baseada (*itDec*), como, por exemplo literatura científica ou

[4] <https://www.gartner.com/en> - empresa criada no final dos anos 1970, com atuação nos ramos de pesquisa, consultoria, eventos e prospecção do mercado de Tecnologia da Informação.

análises de mercado, bem como custos de serviços de consultoria externa (*sCon*). O custo total com as despesas de tempo resultam do tempo total gasto por todos os funcionários $\left(C^{dEst}_{dTempo} = \sum p^{dEst}_{dTempo,m} * a^{dEst}_{dTempo,m}\right)$, ou seja, é determinado pelo valor da hora de trabalho de cada empregado $\left(p^{dEst}_{dTempo,m}\right)$, multiplicado pelo tempo gasto $\left(a^{dEst}_{dTempo,m}\right)$, e somando os valores de todos os empregados (m) envolvidos. Custos com tomadas de decisão ocorrem em períodos $i < 1$ e, além disso, o custo total com a aquisição de informações (*itDec*) pode ser descrito como o somatório do total de preços (p) de todos os materiais adquiridos $\left(C^{dEst}_{itDec} = \sum p^{dEst}_{a}\right)$. Por fim, os custos com serviços de consultoria $\left(C^{dEst}_{sCon}\right)$ são adicionados ao total e todos os gastos correspondentes aos fatores que influenciam o tipo de custo *dEst*, sumarizados pela fórmula $\left(C^{dEst} = C^{dEst}_{dTempo} + C^{dEst}_{sCon} + C^{dEst}_{itDec}\right)$. Para o processo de avaliação e seleção de prestadores de serviços (*ava*), os custos dependem da quantidade de tempo que os empregados dedicam a este processo (*dTempo*) e os custos de eventuais consultorias externas (*sCon*). Os cálculos para $\left(C^{ava}_{dTempo}\right)$ e $\left(C^{ava}_{sCon}\right)$ são análogos a $\left(C^{dEst}_{dTempo}\right)$ e $\left(C^{dEst}_{sCon}\right)$.

Para implementações de nuvem do tipo Infraestrutura como Serviço (IaaS), o tipo de custo que relaciona as taxas sobre os serviços (*txIaaS*) é composto por elementos como o custo com poder de computação (*pComp*), que pode ser calculado multiplicando o número de unidades de processamentos utilizadas $\left(a^{txIaaS}_{pComp,i}\right)$ por período i, pelo custo de uma unidade de processamento $\left(p^{txIaaS}_{pComp,i}\right)$. O preço varia de acordo com as características específicas do sistema, como, memória RAM, número de unidades de computação, capacidade de armazenamento, sistema operacional $\left(C^{txIaaS}_{pComp,i} = a^{txIaaS}_{pComp,i} * p^{txIaaS}_{pComp,i}\right)$ e o custo total deste fator *pComp* resulta do somatório de preços durante todos períodos n $\left(C^{txIaaS}_{pComp,i} = \sum_{i=1}^{n} C^{txIaaS}_{pComp,i}\right)$. Com a generalização do cálculo $\left(C^{t}_{f,i} = \sum_{i=1}^{n} a^{t}_{f,i} * p^{t}_{f,i}\right)$, que sumariza os custos unitários em função da quantidade consumida em um período de uso i, pode-se aplicar o mesmo raciocínio para os valores de capacidade de armazenamento (*cArm*), transferência de dados de entrada (*tEnt*), transferência de dados de saída (*tSda*) e a transferência de dados interna para outros serviços do mesmo provedor (*tInnt*) e o custo com o número de consultas (*nCon*). Os custos com manutenção de domínio para acesso Web (*dom*), certificados SSL (*ssl*), licenciamento de software (*lic*) e taxas básicas de serviços (*txServ*) podem ser determinados pela multiplicação do número de períodos utilizados n pelo respectivo preço p^{t}_{f} do fator de custo f do respectivo tipo de custo t $\left(C^{t}_{f} = n * p^{t}_{f}\right)$.

As despesas com o tempo (*dTempo*) necessário para cumprir as tarefas de implementação, configuração, integração e migração de serviços e dados influenciam o custo total tipo de custo (*imp*). Um fator de custo importante neste tipo de custo é o processo de portabilidade (*pPort*) dos dados do cliente para provedor de serviços. Conforme mencionado, os provedores cobram de seus clientes pela transferência de dados de entrada. Os custos da transferência inicial de dados para a nuvem para fins de migração do sistema pertencem a este tipo de custo. Eles são calculados multiplicando o volume de dados por unidade (ou seja, gigabyte) pelo preço de uma unidade. Alguns provedores oferecem serviços de envio de disco rígido para inserir os dados do cliente. No entanto, essa abordagem não se concentra no volume de dados, mas sim no número de discos rígidos e no tempo de carregamento de dados. O fator de custo de *pPort* não depende de mudanças temporais de preço porque se presume que o processo de portabilidade de dados pode ser concluído dentro de um período t: $\left(C^{imp}_{pPort} = a^{imp}_{pPort} * p^{imp}_{pPort}\right)$. As depesas de tempo C^{imp}_{dTempo} podem ser determinadas da mesma maneira que C^{dEst}_{dTempo}: $\left(C^{imp}_{dTempo} = \sum p^{imp}_{dTempo,m} * a^{imp}_{dTempo,m}\right)$.

O tipo de custo Suporte (*sup*) depende do custo dos serviços de atendimento via telefone, e-mail, sistema de chamados ou *chat* durante todo ciclo de vida da infraestrutura de nuvem. Portanto, este tipo de custo depende do gasto de tempo (*dTempo*) necessário para interação com a equipe de suporte, bem como dos custos ocorridos. Alguns provedores de serviços cobram seus usuários com base no tempo necessário para a resolução de problemas e suporte. Os custos totais com suporte podem ser determinados pela multiplicação do preços de uma unidade pelo número total de unidades necessárias $\left(C^{sup}_{cSup} = p^{sup}_{cSup} * a^{sup}_{cSup}\right)$. Já os custos com resolução de problemas dependem do número de unidades consumidas e o preço por unidade $\left(C^{sup}_{rProb} = p^{sup}_{rProb} * a^{sup}_{rProb}\right)$.

Os custos totais do tipo de custo "treinamento inicial e permanente"(*trein*) podem ser subdivididos em treinamento interno, no qual os próprios colaborados atuam como treinadores, e treinamento externo, no qual são necessários treinadores externos à empresa. Os custos de um treinamento interno dependem da quantidade de tempo de preparação investido por um ou mais empregados (*tPrep*), o tempo de participação dos funcionários internos (*tPart*) e os custos com material para os treinamentos (*mInstr*):

$$\begin{aligned} C^{trein}_{int} &= \sum C^{trein}_{tPrep,m} + \sum C^{trein}_{tPart,m} + C^{trein}_{mInstr,m} \\ &= \sum \left(p^{trein}_{tPrep,m} * a^{trein}_{tPrep,m}\right) \\ &+ \sum \left(p^{trein}_{tPart,m} * a^{trein}_{tPart,m}\right) + C^{trein}_{mInstr} \end{aligned}$$

O total dos custos de um treinamento externo pode se calculado pela adição dos custos com serviços de consultoria que organizam o treinamento (*sCon*), a quantidade de tempo que os empregados investem na participação do treinamento (*tPart*) e os custos com materiais de treinamento (*mInstr*):

$$\begin{aligned} C^{trein}_{ext} &= \sum C^{trein}_{sCon} + \sum C^{trein}_{tPart} + C^{trein}_{mInstr,m} \\ &= \sum \left(p^{trein}_{sCon} * a^{trein}_{sCon}\right) \\ &+ \sum \left(p^{trein}_{tPart,m} * a^{trein}_{tPart,m}\right) + C^{trein}_{mInstr} \end{aligned}$$

Custos com manutenção e modificação (*manut*) dependem das despesas com tempo gasto (*dTempo*) em manutenções ge-

rais e em modificações feitas para implementação de serviços $\left(C^{manut}_{dTempo}\right)$. O cálculo das despesas de tempo para uma respectiva tarefa de manutenção é baseada na fórmula do tipo de custo *dEst*: $\left(c^{manut}_{dTempo} = \sum p^{dEst}_{dTempo,m} * a^{dEst}_{dTempo,m}\right)$.

Custos totais de uma falha de sistema precisam ser declarados para cada empresa individualmente. Os possíveis fatores de custo são, por exemplo, perda de tempo de trabalho produtivo, penalidades contratuais por atrasos ou danos à reputação da empresa, que são difíceis de mensurar. Assim, apenas destaca-se uma fórmula geral que representa a perda por período *i*:

$$C^{falha}_{perda} = \sum_{i=1}^{n} a^{falha}_{perda,i} * p^{falha}_{perda,i}$$

O processo de descarte, ou *backsourcing*, de um sistema envolve despesas de tempo (*dTempo*) e processo de portabilidade (*pPort*). No entanto, os custos com a portabilidade dos dados entre nuvens, ou para um sistema diferente, fazem parte do TCO de um novo serviço e não do TCO do sistema no qual os dados estão sendo retirados. Os custos podem ser determinados da mesma maneira que os custos do processo de portabilidade dos dados para a nuvem $\left(C^{bs}_{pPort} = a^{bs}_{pPort} * p^{bs}_{pPort}\right)$ e também dependem do gasto de tempo necessário para a decisão estratégica necessária:

$$C^{bs}_{dTempo} = \sum p^{bs}_{dTempo,m} * a^{bs}_{dTempo,m}$$

2.3 Custo Global

Em um ambiente de nuvem pública do tipo IaaS, o custo total de propriedade de um serviço de computação em nuvem é igual à soma de todos os tipos de custos envolvidos e pode ser definido como:

$$TCO_{Nuvem} = \sum C^t \text{ onde } t \in T \quad (1)$$

O valor total de um tipo de custo t é igual à soma de todos os fatores de custo f envolvidos, conforme segue: $C^t = \sum C^t_f$ onde $t \in T, f \in F$.

Neste modelo de TCO é considerado o período total de tempo no qual os serviços de nuvem foram ou serão utilizados. Este período é subdividido em vários períodos menores *i*, com duração de um mês, geralmente, predefinido pelo provedor de serviços. Assim, o período total é composto por n períodos menores, de acordo com $C^t_f = \sum^n_i C^t_{f,i}$ onde $i = \{1, \ldots, n\}, t \in T, f \in F$. As variáveis $a^t_{f,i}$ e $p^t_{f,i}$ são utilizadas, respectivamente, para representar a quantidade consumida ou necessária no período i e caracterizar os custos ou preços unitários na fórmula $C^t_{f,i} = a^t_{f,i} * p^t_{f,i}$.

Ao contrário de serviços entregues por meio de nuvens públicas, nas estruturas de nuvens privadas os usuários e provedores fazem parte da mesma organização ou os serviços são prestados, de modo exclusivo, por terceiros. No primeiro cenário, os custos envolvidos incluem, por exemplo, licenciamento de softwares implementados bem como a infraestrutura de TI que deve ser fornecida pela organização. Já no caso de entrega de serviços por provedor exclusivo, o fornecimento é semelhante a uma nuvem pública IaaS, no modo em que o usuário obtém os recursos de um provedor. Apesar disso, o provedor não gerencia dados em uma estrutura pública, mas, sim, em uma nuvem privada exclusiva.

Finalmente, em soluções de nuvens híbridas, que agregam as características de nuvens públicas e privadas, as despesas totais são iguais aos custos totais, ou pelo menos proporcionais, envolvidos em cada solução individual. Além disso, despesas com o processo de agregação de soluções (públicas e privadas) devem ser consideradas na composição dos custos e investimentos.

2.4 Aspectos de Custo em Ambientes HPC

Aplicações para Processamento de Alto Desempenho possuem diferentes características que podem determinar sua adequação a um ambiente de nuvem. Questionamentos a respeito de por que e quem deve escolher uma nuvem para execução de aplicações HPC, quais destas aplicações e como a nuvem pode ser usada para HPC devem balizar os estudos de viabilidade de migração para ambientes remotos distribuídos.

Ambientes HPC são orientados a desempenho, ao passo que nuvens computacionais são orientadas pela relação custo (monetário) *vs.* utilização de recursos. Além disso, nuvens foram, originalmente, projetadas para a execução de aplicações comerciais e de serviços para a internet. O uso de conexões de rede não apropriadas para alto desempenho, a sobrecarga das técnicas de virtualização e as limitações dos sistemas de armazenamento podem ser considerados barreiras para a adoção de nuvens para aplicações HPC. Cabe, então, identificar [10]:

- *Quem* é o usuário candidato para um ambiente de nuvem;
- *Qual* é o tipo de aplicação que pode se beneficiar de uma execução em ambiente de nuvem;
- *Por que* o usuário terá benefícios na execução de sua aplicação em ambiente de nuvem; e,
- *Como* este benefício pode ser atingido.

Com base nestes pontos, as tabelas 2 e 3 apresentam alguns posicionamentos com relação aos questionamentos originados a partir de duas diferentes abordagens que visam auxiliar nas decisões de migração para ambientes de nuvem computacional: *a*) considera-se os aspectos da execução em modelos de desempenho, custo e negócios; e *b*) são exploradas técnicas para preencher as lacunas entre nuvens e aplicações HPC. Estas abordagens têm por objetivo identificar as diferenças de demandas por HPC, por parte dos usuários, e quais alternativas eles dispõem para implantar suas aplicações. As alternativas que buscam preencher a lacuna entre aplicações HPC e nuvens podem ser classificadas em duas categorias [10]: primeiro, aquelas que objetivam tornar as nuvens cientes das aplicações HPC e, segundo, aquelas que buscam tornar as aplicações HPC cientes das infraestruturas de nuvem nas quais serão executadas.

A Tabela 2 traz respostas para a o cenário no qual é pretendido tornar as nuvens cientes das aplicações HPC que serão executadas. Para isso, são exploradas técnicas de virtualização leve (por exemplo, contêineres) e determina-se quão próximo do desempenho de máquina física é possível chegar com as aplicações HPC. São exemplos, *hypervisors* otimizados para HPC e nuvens otimizadas para computação de alto desempenho como Amazon HPC.

Tabela 2. Questões sobre nuvens cientes de aplicações HPC

Questionamento	Resposta
Quem	Pequenas e médias organizações ou empresas em crescimento.
Qual	Aplicações com padrões de comunicação menos intensos e menos sensíveis a interferências.
Por que	Pequenas e médias organizações que são sensíveis aos argumentos de CAPEX/OPEX.
Como	Tornar as nuvens cientes das aplicações HPC, por exemplo, com uso de virtualização leve e afinidade de CPU.

Fonte: Adaptado de [10].

Já na Tabela 3, as respostas tratam de pontos que devem ser considerados quando se busca por ambientes de execução capazes de trabalhar com aplicações HPC cientes de nuvens computacionais. Esta alternativa, embora ainda pouco explorada, permite ajustes nos tempos de execução das aplicações HPC em nuvens para obtenção de um melhor desempenho com uso de tecnologias como balanceadores de carga com reconhecimento de nuvem para aplicações HPC e a implantação de topologias com reconhecimento de aplicações científicas na nuvem.

Tabela 3. Questões sobre aplicações HPC cientes de nuvem

Questionamento	Resposta
Quem	Usuários com aplicações que tem melhor relação custo/-desempenho em nuvens *vs.* outras plataformas.
Qual	Aplicações com necessidades de desempenho que podem ser atendidas em média escala (em termos de número de *cores*).
Por que	Nuvens permitem que vários usuários acessem estruturas compartilhadas, garantindo um melhor uso dos recursos.
Como	Abordagem híbrida supercomputador-nuvem com escalonamento ciente de aplicação e *cloud bursting*.

Fonte: Adaptado de [10].

O uso de nuvens para execução de aplicações HPC pode ser visto como um bom complemento para estruturas locais de supercomputadores e *clusters*, não podendo, ainda, substituí-las totalmente. Abordagens de utilização de nuvens híbridas, nas quais ocorre uma integração entre infraestruturas de nuvens públicas e privadas, permitem a ocorrência de *cloud bursting* [11]. Isto faz com que aplicativos utilizem toda a capacidade dos recursos computacionais de uma nuvem privada e, em reação a este aumento, migurem suas tarefas em ambientes de nuvem pública, à medida que os recursos locais se tornem escassos.

Aplicações científicas tem necessidades significativamente diferentes das aplicações comerciais típicas, executando suas tarefas de maneira fortemente acoplada e em escalas bem maiores de recursos computacionais. Tal comportamento leva a requisitos de largura de banda e latência mais exigentes do que a maioria dos usuários de nuvem. Aplicações científicas também requerem acesso a grandes quantidades de dados e isso pode levar a um grande custo de inicialização e armazenamento [12].

Cargas de trabalho científicas podem ser classificadas em três categorias abrangentes de acordo com seus requisitos: fortemente acopladas em grande escala, médio alcance e alto rendimento. A Tabela 4 apresenta características que favorecem o entendimento da viabilidade de execução de aplicações HPC em nuvens. Também classifica, em alto nível, cargas de trabalho executadas pela comunidade científica.

Tabela 4. Recomendação de uso de nuvens em função do tipo de aplicação HPC

Tipo de Aplicação	Recomendação
Fortemente acoplada em grande escala	Tipicamente, aplicações MPI, que utilizam milhares de *cores* e exigem uma rede de comunicação de alto desempenho. Neste caso, qualquer gargalo de virtualização ou problemas de alta latência na rede terão impacto negativo no desempenho das aplicações. Logo, é recomendável a execução em ambientes tradicionais de supercomputação ou em nuvens privadas com servidores *bare metal* e redes de alta velocidade.
Médio alcance	Estas aplicações utilizam um número variado de *cores* (dezenas ou centenas) e têm requisitos de desempenho mais baixos do que as aplicações de grande escala fortemente acopladas. Consequentemente, são mais tolerantes à virtualização e redes tradicionais. É recomendado que estas aplicações explorem os benefícios do acesso rápido a recursos para a nuvem, especialmente a virtualização leve (contêineres).
Alto rendimento	Aplicações compostas por tarefas independentes que exigem pouca ou nenhuma comunicação. Tais aplicações podem se beneficiar de um grande número de recursos disponíveis e são tolerantes à heterogeneidade. Recomenda-se o uso de nuvens para estas aplicações, especialmente explorando mecanismos de elasticidade. Outros benefícios também podem ser alcançados por meio de ambientes de nuvem híbrida HPC, distribuindo tarefas em *clusters* locais e nuvens públicas.

Fonte: Adaptado de [12].

A partir da classificação das cargas de trabalho científicas, caracterizadas anteriormente (Tabela 4), é possível identificar alguns exemplos de cada um dos tipos de aplicação, conforme descrito na Tabela 5.

Em seu estudo, o grupo de Parashar [13] apresentou uma divisão dos ambientes de nuvem para HPC em três categorias: *a*) ***HPC in the Cloud***, que se concentra em mover aplicações HPC para ambientes de nuvem; *b*) ***HPC Plus Cloud***, na qual usuários fazem uso de nuvens para complementar seus recursos de HPC, em situações de picos de demanda (*cloud bursting*); e *c*) ***HPC as a Service***, que expõe os recursos HPC por meio de serviços de nuvem. Estas categorias estão rela-

Tabela 5. Exemplos de aplicações HPC

Tipo de Aplicação	Exemplos
Fortemente acoplada em grande escala	Aplicações MPI; aplicações de geração de modelos climáticos, sísmicos.
Médio alcance	Aplicações de simulação orientada a eventos; aplicações escalonadas no tempo, cujos trabalhos são menos sensíveis a prazos.
Alto rendimento	Aplicações BoT; aplicações MapReduce; simulações Monte Carlo.

cionadas a como os recursos são alocados e abstraídos para simplificar o uso da nuvem.

A execução de aplicações HPC em nuvem ainda possui vários problemas em aberto. Para exemplificar, a abstração da infraestrutura de nuvem limita o ajuste das aplicações. Além disso, a maioria das conexões de rede dos provedores de nuvem não é rápida o suficiente para aplicações fortemente acopladas em grande escala, com alta comunicação entre processadores.

O modelo de negócio para uma nuvem HPC também é um campo a ser bem explorado. Em nuvens públicas, provedores de serviços lançam várias cargas de trabalho sobre os mesmos recursos físicos a fim de explorar economia de escala, que nem sempre é apropriada para HPC. Além disso, mesmo que pequenas empresas se beneficiem do rápido acesso a recursos de nuvens públicas, isso nem sempre é verdadeiro para grandes usuários de HPC [4].

Por outro lado, em ambientes de nuvens privadas, usuários de HPC podem ter acesso direto ao gerenciamento dos componentes e o compartilhamento de recursos é reduzido por consequência de o modelo de múltiplos inquilinos ficar limitado a um grupo específico de usuários.

Questões referentes à troca de CAPEX por OPEX tem destaque no processo de decisão de migração de aplicações para ambientes de nuvem. CAPEX está relacionado com os investimentos feitos na aquisição de equipamentos, softwares e inicialização dos mesmos dentro do provedor de serviço. Já o OPEX diz respeito aos investimentos feitos com alocação de serviços, como por exemplo, manutenção de equipamentos e contratação de serviços de nuvem.

Aplicações que fazem uso variável dos recursos computacionais determinam uma menor utilização global dos equipamentos. Juntamente com os pontos referentes a CAPEX e OPEX, esta utilização variável de recursos serve de argumentos para provedores e usuários finais de nuvem. Os usuários podem se beneficiar caso suas execuções se caracterizem como aplicações de uso variável [10]. Por outro lado, provedores de nuvem podem tirar benefícios de uma utilização agregada de recursos de todos os seus inquilinos. Para isso, é fundamental que a agregação possa sustentar um modelo de precificação lucrativo frente aos grandes investimentos iniciais necessários para oferecer recursos de computação e armazenamento por meio de uma interface em nuvem pública. No caso de uma nuvem privada, o lucro está em oferecer mais produtividade para os usuários, atendendo de modo satisfatório às demandas da instituição.

2.5 Discussão

Aplicações HPC necessitam de muitos recursos computacionais e fornecê-los de maneira otimizada requer ajustes em vários aspectos [14]. Em termos de desempenho, o uso de virtualização já está bem mais aprimorado, devido ao suporte à virtualização no nível de sistema operacional, fornecendo desempenho próximo às infraestruturas de *bare-metal*.

Do ponto de vista do fluxo de trabalho, a transição de uma infraestrutura de hardware dedicado para serviços oferecidos em nuvem implica em uma modificação de processos. Isto é válido para qualquer tipo de aplicação, inclusive para as aplicações HPC. Esta mudança de processo é especialmente complicada devido ao armazenamento e à transferência de grandes quantidades de dados. Este aspecto pode ser simplificado com a hospedagem dos dados diretamente no provedor de serviços.

Os ambientes de nuvem e as estruturas clássicas para HPC têm maneiras distintas de gerenciar recursos de computação. Para extrair o melhor desempenho de um ambiente em nuvem, os esforços concentraram-se em aumentar o isolamento entre máquinas virtuais e reduzir a sobrecarga imposta pelas técnicas de virtualização. As estruturas HPC, por sua vez, visam extrair o máximo de desempenho possível da infraestrutura [4].

As solicitações de usuários para acessar exclusivamente partes de um *cluster* HPC são enfileiradas sempre que os recursos estiverem sobrecarregados. Em ambientes de nuvem isso não ocorre devido à disponibilidade "ilimitada" de recursos, também chamada de elasticidade. Além do mais, hardware para HPC, especificamente interfaces de rede, são consideravelmente mais caras do que exemplares utilizados para criação de nuvens tradicionais, com hardware de *commodity*[5]. Sendo assim, alocar usuários em hardwares específicos para HPC, configura um desperdício para o provedor de serviços, caso ele não atenda exclusivamente usuários com demandas HPC.

O desafio em gerenciar recursos HPC em nuvens está em desenvolver um modelo de negócio sustentável, com economia de escala e que ofereça processamento de alto desempenho aos usuários. Para avançar nesta área, esforços de pesquisa devem ser concentrados para que os provedores possam oferecer sistemas de filas e modelos de preços que levem em conta o tempo que os usuários estão dispostos a esperar para ter acesso aos recursos. Alguns modelos flexíveis de aluguel de recursos HPC em nuvens, como apresentado em [15], são baseados em estratégias de planejamento que consideram instâncias sob demanda e *spot*, motivados pelos interesses das duas partes, usuário e provedor, envolvidas no processo de alocar recursos.

[5]Em um contexto de Tecnologia da Informação, é um dispositivo ou componente de dispositivo que é relativamente barato, amplamente disponível e, em alguns casos, intercambiável com outro hardware do mesmo tipo.

3. Revisão Sistemática

No processo de pesquisa e seleção dos trabalhos relacionados ao tema deste estudo, foi realizada uma Revisão Sistemática da Literatura (RSL), desenvolvendo seus três grandes estágios: planejamento, condução e relatório da revisão [16]. Na fase de planejamento, é identificada a necessidade de uma revisão, com a especificação de questões de pesquisa e desenvolvimento de um protocolo de revisão. Na condução da revisão, são identificados e selecionados os estudos primários, extraídos os dados, analisados e sintetizados. Por fim, no relatório da revisão sistemática, são divulgadas as descobertas e discutidos os trabalhos restantes da pesquisa.

3.1 Planejamento da RSL

Para o desenvolvimento da RSL no tema proposto, durante o estágio de planejamento, foram definidas algumas Questões de Pesquisa (QPs). As QPs foram elaboradas a partir de estudos preliminares e conversas com especialistas no tema, membros do mesmo grupo de pesquisa dos autores, com o objetivo de nortear e fundamentar o estudo. As Questões de Pesquisa são:

QP 1: Qual o impacto econômico nas decisões de implantação de aplicações de alto desempenho em nuvens computacionais?

Esta QP busca identificar como instituições (acadêmicas ou comerciais) podem se beneficiar do uso de nuvens computacionais para execução de aplicações que demandem alto poder de processamento. Esta questão considera os custos financeiros envolvidos.

QP 2: Como os usuários podem identificar qual a melhor configuração de nuvem, seja pública ou privada, para executar suas aplicações?

Nesta questão, o objetivo é prover subsídios para que os usuários de nuvem possam determinar qual opção de configuração, oferecida pelos provedores de serviço, mais se adapta às necessidades de suas demandas de processamento. Neste aspecto deve ser levado em conta que existe perda financeira à medida que houver super ou subdimensionamento de recursos.

QP 3: Quais modelos de custo financeiro são empregados na execução de aplicações de alto desempenho em nuvens?

Com esta questão, busca-se identificar possíveis modelos de custo utilizados por instituições para definir as demandas de infraestruturas de nuvem para execuções de aplicações com demanda de processamento de alto desempenho.

QP 4: Como identificar eventuais perdas financeiras em função de imprecisão no dimensionamento de infraestruturas de nuvens para aplicações com demandas de HPC?

O objetivo desta questão é identificar se haverá perda financeira em função do dimensionamento inadequado da infraestrutura para a demanda do usuário. Neste caso, pode ocorrer superdimensionamento, sendo dispendidos mais recursos na infraestrutura que o necessário para atender à demanda, ou subdimensionamento, quando a execução da aplicação pode atrasar devido à insuficiência de recursos computacionais para atender à demanda. No primeiro caso, a perda está relacionada ao passivo em recursos instalados, e as consequentes despesas em manutenção. No segundo, a perda está relacionada ao baixo índice de produção e a consequente perda de lucros.

De posse destas Questões de Pesquisa, foi realizada uma busca exploratória à procura de trabalhos que abordassem os principais temas contidos nas perguntas. Com isso foi possível identificar como a comunidade científica faz referência aos temas e também extrair palavras-chave utilizadas para composição da *string* de busca utilizada na presente RSL, conforme documentado na sequência (Seção 3.2).

3.2 Condução da RSL

Para a condução desta RSL foram acessadas diversas bases de indexação de trabalhos, amplamente utilizadas por pesquisadores. Ao todo foram selecionadas cinco bases: *i*) ACM Digital Library; *ii*) IEEE Digital Library; *iii*) Science@Direct; *iv*) Scopus; e *v*) Web of Science. Tais bases foram escolhidas devido a sua importância e por serem repositórios digitais que oferecem acesso eletrônico à maioria dos periódicos e artigos de conferências publicados na área da Ciência da Computação [17].

Em seguida, um conjunto de termos de pesquisa foi identificado para compor a *string* de busca com a qual será possível extrair da literatura trabalhos relacionados com o tema abordado. Os termos selecionados visam identificar, na literatura, trabalhos envolvendo ambientes de nuvens (*((cloud OR "cloud computing"))* e computação de alto desempenho (*((hpc OR "high performance computing"))*. A esta *string* foram associados termos referentes a possíveis modelos de custo e preços praticados em situações que contemplem a execução de aplicações de HPC em ambientes de nuvem, analisando sua viabilidade financeira. Em uma etapa de identificação de termos relevantes para construção da *string* de busca, foram utilizados termos como *cost*, *pricing*, *investiment* e *return*. Nesta etapa preliminar, a análise dos artigos, e de suas palavras-chave permitiu identificar como relevante os termos: (*("cost model" OR "cost efficiency" OR "cost analysis" OR "economic analysis" OR "monetary cost" OR "billing model" OR "price efficiency" OR investment OR pricing OR price)*). A *string* de busca concebida é:

((cloud OR "cloud computing") AND (hpc OR "high performance computing") AND ("cost model" OR "cost efficiency" OR "cost analysis" OR "economic analysis" OR "monetary cost" OR "billing model" OR "price efficiency" OR investment OR pricing OR price))

De posse da *string* de busca, foram realizadas pesquisas nas bases selecionadas. Uma das características disponíveis nas bases de indexação utilizadas é a possibilidade de exportação dos resultados para o formato *BibTeX*[6]. Os resultados alcançados por meio da execução de buscas nas bases foram

[6]Ferramenta para formatação de bibliografias utilizada em documentos LaTeX.

extraídos para arquivos (*.bib*) e, posteriormente, importados na ferramenta Parsifal[7] que auxiliará na análise dos resultados, dando prosseguimento ao processo de revisão.

O processo de busca por artigos desta RSL se desenvolveu em quatro fases, cada uma com critérios de exclusão associados. A primeira fase consiste na aplicação da *string* de busca nas bases de indexação. Na sequência, em cada uma das demais fases, serão aplicados filtros nos resultados em conformidade com os critérios de exclusão da Tabela 6.

Tabela 6. Critérios de exclusão

Fase	ID	Critério de exclusão
Fase 1	-	Aplicação da *string* de busca nas bases de indexação.
Fase 2	2.1	Trabalhos anteriores ao ano de 2010.
Fase 3	3.1	Trabalhos que não contêm a *string* de busca no título, resumo ou palavras-chave.
Fase 4	4.1	Trabalhos duplicados.
	4.2	Trabalhos que não estão publicados em conferências ou periódicos.
	4.3	Trabalhos nos quais título e resumo não abordam o tema de estudo.
	4.4	Trabalhos que não apresentem uma avaliação de custos de infraestrutura.
	4.5	Trabalhos que não relacionem nuvens e execução de aplicações HPC.
	4.6	Trabalhos sem acesso ao texto completo.
	4.7	Trabalhos não classificados pela avaliação de qualidade.
	4.8	Trabalhos que apresentem pequenas modificações de estudos do mesmo grupo.

Na segunda fase é aplicado o critério 2.1, na qual fica explícito que trabalhos anteriores ao ano de 2010 encontram-se desatualizados ou, caso tenham sido continuados, novos resultados devem estar contemplados em publicações mais atuais.

Na terceira fase da revisão foi aplicado o critério de exclusão 3.1, retirando os trabalhos que não possuem os termos da *string* de busca no título, resumo ou palavras-chave. Este critério visa reduzir o número de resultados falso-positivos da pesquisa realizada na busca inicial (fase 1).

Já na quarta e última fase, os demais critérios de exclusão são aplicados. Esta fase requer uma análise mais detalhada dos textos, com base nos critérios restantes. O objetivo da fase quatro é selecionar somente os trabalhos que abordam o tema de pesquisa relacionado, apresentando avaliações de custos de infraestruturas locais e de nuvem, bem como relacionando a execução de aplicações com demandas de alto desempenho em ambientes de nuvem.

Na fase 4, o critério 4.1 realiza a busca por trabalhos duplicados, removendo-os da seleção, enquanto o critério 4.2 procura por estudos que não estão publicados em conferências ou periódicos. Para o critério de exclusão 4.3, foi realizada a leitura dos títulos e resumos a fim de retirar artigos que não abordam o tema de pesquisa objeto deste estudo.

Neste ponto, para aplicação dos critérios 4.4 e 4.5, que buscam, respectivamente, por trabalhos que não apresentam uma avaliação de custos de infraestrutura e não relacionam nuvens com a execução de aplicações HPC, foi necessária uma leitura completa dos trabalhos. Durante o processo de leitura, não foi possível obter acesso aos textos completos de alguns trabalhos que, por consequência, foram excluídos pelo critério 4.6.

Para avaliação dos trabalhos selecionados até este momento da RSL, são aplicadas algumas questões de qualidade (critério de exclusão 4.7), conforme segue:

- Os objetivos da pesquisa estão claramente especificados?
- O trabalho considera a satisfação do usuário?
- O modelo de custo financeiro considera o desempenho das aplicações?
- O trabalho considera a satisfação do provedor de serviços?
- O desempenho da execução de aplicações é considerado?
- O trabalho apresenta resultados com relação aos custos?

Cada questão de qualidade possui três opções de respostas: “sim”, “parcialmente” ou “não”, com valores atribuídos, respectivamente, “1”, “0.5” e “0”. Os trabalhos podem ser pontuados com, no máximo, seis pontos e no mínimo, zero. Foi definida a pontuação “3.5” como ponto de corte, ou seja, pontos mínimos para ser considerado aceito.

Ao final da fase 4, o último critério de exclusão, 4.8, foi aplicado aos textos desta RSL com o objetivo de identificar trabalhos que tragam pequenas modificações de trabalhos anteriores de um mesmo grupo de pesquisa.

A Tabela 7 apresenta, de modo geral, os quantitativos de trabalhos suprimidos em cada fase da revisão, de acordo com os critérios de exclusão definidos. Conforme os valores demonstrados na tabela, percebe-se que o critério que mais excluiu trabalhos pertence à fase 3, sendo o 3.1, o qual exclui trabalhos que não contêm a *string* de busca no título, *abstract* ou palavras-chave dos materiais analisados, totalizando 4593 trabalhos.

Tabela 7. Quantitativo de trabalhos rejeitados em cada critério de exclusão

Base	Critérios de exclusão										
	F1	F2	F3	F4							
	-	2.1	3.1	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8
ACM Digital Library	134	3	62	47	9	8	0	1	1	3	0
IEEE Digital Library	509	11	367	48	2	62	6	9	0	2	0
Science@Direct	328	0	323	2	0	1	1	0	0	1	0
Scopus	3918	6	3761	23	19	58	14	4	4	12	7
Web of Science	168	1	80	76	1	8	2	0	0	0	0
Total	5057	21	4593	196	31	137	23	14	5	18	7
Trabalhos restantes	5057	5036	443	247	216	79	56	42	37	19	12

Por fim, esta RSL selecionou 12 trabalhos, apresentadas na Tabela 8, que versam sobre a análise de investimentos em infraestruturas de nuvem, levando em consideração modelos

[7]Ferramenta Web para Revisões Sistemáticas de Literatura. Disponível em <https://parsif.al/>

de custo para execução de aplicações HPC. A Seção 3.3 trará um apanhado sobre os artigos, apresentando-os com maiores detalhes.

3.3 Relatório da RSL

O último estágio desta RSL traz a etapa de extração de dados dos trabalhos selecionados pela revisão. Tal processo se dará com a análise dos artigos, buscando recuperar os conteúdos referentes ao problema abordado, solução proposta, método utilizado para validação da proposta e trabalhos futuros. Na sequência, os dados extraídos de cada trabalho são apresentados.

- **High Performance Computing in the cloud: Deployment, performance and cost efficiency [2]**

  Problema: A execução de aplicações de alto desempenho (HPC) em nuvem alcançou um lugar de destaque e se tornou um importante tópico de pesquisa científica e para a indústria. No entanto, o foco dos provedores de serviços em nuvem é a entrega de conteúdo e não HPC. Por esse motivo, faltam pesquisas direcionadas à implantação de aplicações HPC em infraestruturas de nuvem.
  Proposta: Realizar uma avaliação abrangente de três aspectos importantes da execução de aplicações HPC em nuvem: implantação, desempenho e eficiência de custos. Para a avaliação foram utilizados conjuntos de *benchmarks* conhecidos, como o *NAS Parallel Benchmarks* (NPB), e as execuções ocorreram em três provedores de nuvem diferentes: Amazon Elastic Cloud Compute (EC2), Microsoft Windows Azure e Rackspace.
  Validação: Os resultados das avaliações foram comparados com um *cluster* real, com características semelhantes às instâncias das nuvens utilizadas, para analisar diferenças de eficácia de custos e desempenho e, também, prover argumentos para discussões sobre quando aplicações de alto desempenho fazem sentido em ambientes de nuvem e para quais casos de uso.
  Trabalhos Futuros: Migrar aplicações HPC completas para ambientes de nuvem e estender a métrica de custo e eficiência para cobrir mais fatores e ser mais flexível.

- **A comparative study of high-performance computing on the cloud [18]**

  Problema: A popularidade da plataforma em nuvem EC2 da Amazon aumentou nos últimos anos. No entanto, muitos usuários de computação de alto desempenho (HPC) consideram que os *clusters* dedicados de alto desempenho, normalmente encontrados em grandes centros de computação, são muito superiores ao EC2, devido à significativa sobrecarga de comunicação deste último.
  Proposta: Examinar, de forma inovadora, diferenças ao comparar os *clusters* Amazon EC2 de ponta com os *clusters* HPC tradicionais, mas com um esquema de avaliação mais geral. Primeiro, comparar o EC2 a cinco *clusters* do *Lawrence Livermore National Laboratory* (LLNL) com base no tempo total de resposta para um conjunto típico de *benchmarks* de HPC em diferentes escalas; para o tempo de espera na fila nos *clusters* HPC, usou-se uma distribuição desenvolvida a partir de simulações com rastreamentos reais. Segundo, para permitir uma comparação do custo total de execução, foi desenvolvido um modelo econômico para precificar *clusters* do LLNL, supondo que eles sejam oferecidos como recursos de nuvem a preços por hora do nó.
  Validação: Para estimar o custo foi desenvolvido um modelo de preços, relativo aos preços por hora por nó do EC2, para que fosse possível estimar os mesmos valores para o *cluster* do LLNL.
  Trabalhos Futuros: Utilizar tempo de resposta e custo de para desenvolver ferramentas e técnicas que direcionam os usuários de diversos conjuntos de aplicativos, dadas restrições específicas, ao *cluster* mais apropriado.

- **A performance/cost model for a CUDA drug discovery application on physical and public cloud infrastructures [19]**

  Problema: Na pesquisa clínica, é crucial determinar a segurança e a eficácia dos medicamentos atuais e acelerar os achados na pesquisa básica, como a descoberta de novos compostos ativos, em resultados significativos para a saúde. Ambos os objetivos requerem o processamento de grandes conjuntos de dados de estruturas de proteínas disponíveis em bancos de dados biológicos.
  Proposta: É proposto um modelo de desempenho/-custo para a aplicação BINDSURF, permitindo que o usuário decida qual infraestrutura, local ou de um conhecido provedor de nuvem pública, é ideal para um determinado tipo e tamanho de problema. A execução de uma aplicação com uso intensivo de GPU, como o BINDSURF, pode sobrecarregar o orçamento de uma instituição ao processar grandes quantidades de dados. Quanto maior o número de recursos físicos computacionais ou maior o tempo de execução para esses recursos, mais o custo total é aumentado, mesmo para infraestruturas locais não utilizadas.
  Validação: Criação de um modelo de custo e desempenho para bioinformática utilizando o algoritmo BINDSURF, para obter o melhor desempenho de execução durante o tempo e otimizar os custos.
  Trabalhos Futuros: Portar BINDSURF para OpenCL, permitindo que ele seja executado em uma variedade maior de sistemas computacionais heterogêneos, como CPUs com vários núcleos. Isso permitirá que um número maior e mais barato de tipos de instâncias de provedores de nuvem pública sejam usados para um modelo mais abrangente de desempenho e custo.

- **Cost-Optimized Resource Provision for Cloud Ap-**

Tabela 8. Trabalhos selecionados na RSL

ID	Título	Citação
T01	High Performance Computing in the cloud: Deployment, performance and cost efficiency	[2]
T02	A comparative study of high-performance computing on the cloud	[18]
T03	A performance/cost model for a CUDA drug discovery application on physical and public cloud infrastructures	[19]
T04	Cost-Optimized Resource Provision for Cloud Applications	[20]
T05	Evaluating and Improving the Performance and Scheduling of HPC Applications in Cloud	[10]
T06	Price efficiency in High Performance Computing on Amazon Elastic Compute Cloud provider in Compute Optimize packages	[21]
T07	Scheduling deadline constrained scientific workflows on dynamically provisioned cloud resources	[22]
T08	Cost Analysis Comparing HPC Public Versus Private Cloud Computing	[23]
T09	HPC Application Performance and Cost Efficiency in the Cloud	[24]
T10	Understanding the Performance and Potential of Cloud Computing for Scientific Applications	[25]
T11	Amazon Elastic Compute Cloud (EC2) versus In-House HPC Platform: A Cost Analysis	[26]
T12	Exploring Instance Heterogeneity in Public Cloud Providers for HPC Applications	[27]

plications [20]

Problema: Com um número crescente de provedores de serviços em nuvem oferecendo locação de recursos virtuais, os provedores de aplicações têm mais opções quando precisam de recursos. Mas alcançar a solução de recursos mais otimizada em custo ainda é um desafio.
Proposta: Uma abordagem para auxiliar os usuários a calcular a quantidade otimizada de recursos virtuais com base na carga de trabalho prevista e resolver soluções de fornecimento de recursos, o que inclui instâncias de máquinas virtuais de diferentes tipos e preços. Os SLAs publicados pelos provedores são atendidos o máximo possível.
Validação: Para estimar a relação entre a carga de trabalho prevista e a quantidade de máquinas virtuais, foram realizados experimentos de projeção, simulando as instâncias do tipo micro do Amazon EC2. Utilizando as políticas de preços do EC2, foi obtida uma solução de provisão de economia de custos para fornecedores de aplicações.
Trabalhos Futuros: Implementar a abordagem proposta como uma estrutura de tempo de execução dinâmico para aplicações em nuvem. Isso envolverá mais modelos de previsão e técnicas de aprendizado de máquina. Além disso, aprimorar o algoritmo de provisionamento para suportar políticas de preços de mais provedores de nuvem existentes.

- **Evaluating and Improving the Performance and Scheduling of HPC Applications in Cloud [10]**

  Problema: A computação em nuvem está surgindo como uma alternativa promissora aos supercomputadores para algumas aplicações de computação de alto desempenho (HPC). Com a nuvem como uma opção de implantação adicional, os usuários e fornecedores de HPC enfrentam os desafios de lidar com recursos altamente heterogêneos, onde a variabilidade se estende por uma ampla gama de configurações de processadores, interconexões, ambientes de virtualização e modelos de preços.
  Proposta: Uma avaliação detalhada e abrangente de desempenho e custo de execução de um conjunto de aplicações HPC em uma variedade de plataformas, variando desde supercomputadores às nuvens computacionais. Este estudo permite uma visão holística que busca responder ao questionamento por que e quem deve escolher uma nuvem para execução de aplicações HPC e quais aplicações e como as nuvens devem ser utilizadas para HPC. Investigar os aspectos econômicos da execução em nuvem e discutir por que é desafiador ou gratificante para os provedores de nuvem operar negócios para a HPC em comparação com as aplicações em nuvem tradicionais.
  Validação: Avaliação de desempenho e gargalos de aplicações HPC em estruturas de supercomputadores, *clusters* e nuvens (privadas e públicas). Com o uso de *benchmarks* executados no mesmo hardware, com e sem uso de *hypervisors*, foi possível uma análise detalhada do impacto do uso de virtualização para HPC. Para a tarefa de investigar a co-existência de várias plataformas foi utilizado o simulador CloudSim.
  Trabalhos Futuros: Considerar outros fatores no escalonamento de várias plataformas: qualidade de serviço (QoS), prazos, prioridades e segurança. Além disso, pesquisas futuras são necessárias no que diz respeito ao preço das nuvens em ambientes de várias plataformas. Outro tema promissor é a avaliação e caracterização de aplicações com paralelismo irregular e conjuntos de dados dinâmicos.

- **Price efficiency in High Performance Computing on Amazon Elastic Compute Cloud provider in Compute Optimize packages [21]**

  Problema: Atualmente, a computação de alto desempenho (HPC) é usada em muitas pesquisas e trabalha para calcular ou processar dados. Neste sentido, em

relação à computação de alto desempenho (HPC), o trabalho pretende dar subsídios aos usuários para que possa ser determinado qual o pacote otimizado de serviços, oferecido pelo provedor, é o mais adequado para uma dada aplicação HPC.
Proposta: Investigar a eficiência de preços que pode beneficiar o cliente na escolha do pacote de otimizações oferecido pelo fornecedor de serviços de nuvem. Analisar a relação entre preços, tempos de execução e tamanhos de problemas ou cargas de trabalho do HPL no pacote otimizado para computação do Amazon EC2.
Validação: Avaliação de todas as instâncias do pacote otimizado para computação do Amazon EC2, usando o *benchmark* HPL, com base na carga de trabalho ou no tamanho do problema de entrada. Para eficiência de preço, a relação de tamanho do problema, tempo em HPL e preço foi analisada para obter uma instância adequada para o uso da computação de alto desempenho.
Trabalhos Futuros: Não apresenta trabalhos futuros.

- **Scheduling deadline constrained scientific workflows on dynamically provisioned cloud resources [22]**

  Problema: Uma nuvem permite que pesquisadores e instituições provisionem recursos de computação apenas quando necessário e escalem conforme necessário. No entanto, ainda existem obstáculos técnicos significativos associados à obtenção de desempenho de execução suficiente e limitação do custo financeiro. Os esforços se concentram no problema de agendamento de cargas de trabalho científicas com restrições de prazo em recursos de nuvem provisionados dinamicamente, enquanto reduz o custo da computação.
  Proposta: São apresentados dois algoritmos, o *Proportional Deadline Constrained* (PDC) e o *Deadline Constrained Critical Path* (DCCP) que abordam o problema de agendamento de fluxo de trabalho nos recursos de nuvem provisionados dinamicamente. Esses algoritmos são adicionalmente estendidos para refinar sua operação na priorização de tarefas e preenchimento, respectivamente.
  Validação: Os algoritmos foram avaliados, por meio de simulação, com o uso do CloudSim, que apresenta recursos de nuvem provisionados dinamicamente e um modelo de pagamento por uso derivado do modelo de precificação EC2 da Amazon. As simulações foram realizadas usando cinco fluxos de trabalho científicos: Montage, SIPHT, LIGO, Cybershake e Epigenomics. Cada fluxo consistiu em 1000 tarefas e foram obtidos a partir do gerador de fluxos de trabalho Pegasus.
  Trabalhos Futuros: Investigar o impacto da estrutura de fluxo de trabalho, procurando uma medida de simetria a fim de considerar como isso pode ser incorporado nas decisões de escalonamento.

- **Cost Analysis Comparing HPC Public Versus Private Cloud Computing [23]**

  Problema: Nos últimos anos, houve um rápido aumento no número e tipo de configurações de hardware de computação em nuvem pública e opções de preços oferecidas aos clientes. Além disso, os provedores de nuvem pública também expandiram o número e o tipo de opções de armazenamento e estabeleceram preços incrementais para armazenamento e transmissão em rede de dados de saída da instalação em nuvem. Tal cenário prejudica a análise para determinar a opção mais econômica em uma migração de aplicações de uso geral para a nuvem.
  Proposta: Investigar se a análise econômica para mover aplicações de uso geral para uma nuvem pública pode ser estendida para execuções do tipo HPC com uso intensivo de computação.
  Validação: Uma comparação de custos com uma determinada configuração de hardware HPC é estabelecida para determinar sob quais condições uma nuvem pública e não uma nuvem privada será mais econômica em cálculos, armazenamento e transferências de dados de rede para aplicativos do tipo HPC.
  Trabalhos Futuros: Não apresenta trabalhos futuros.

- **HPC Application Performance and Cost Efficiency in the Cloud [24]**

  Problema: Nos últimos anos, várias abordagens foram introduzidas para o uso eficiente de nuvens para a execução de aplicações HPC. No entanto, faltam pesquisas sobre oportunidades e desvantagens do uso de nuvens públicas como ambiente eficiente para HPC.
  Proposta: Identificar as instâncias de máquina virtual em nuvens públicas disponíveis que possam ser adequadas para aplicações HPC, tanto em termos de desempenho quanto de eficiência de custos. Também avaliar que tipo de aplicação pode se beneficiar da execução na nuvem. Para isso, é preciso analisar as características das instâncias, levando em consideração os aspectos relevantes para HPC. Também é necessária uma análise da eficiência de custos usando os *benchmarks* tradicionais de HPC.
  Validação: Foi realizada uma extensa avaliação dos dois maiores provedores de computação em nuvem, Amazon EC2 e Microsoft Azure, considerando comunicação de rede, processamento e desempenho de memória.
  Trabalhos Futuros: Adicionar métricas de desempenho para dispositivos de entrada e saída à avaliação, pois tem sido uma área com bastante desenvolvimento na nuvem nos últimos anos. Também avaliar ambientes de nuvem que possuem aceleradores, como GPUs.

- **Understanding the Performance and Potential of Cloud Computing for Scientific Applications [25]**

  Problema: As aplicações científicas geralmente exigem recursos significativos, no entanto, nem todos os

cientistas têm acesso a sistemas de computação de ponta suficientes. A computação em nuvem chamou a atenção dos cientistas como um recurso competitivo para execução de aplicações HPC, a um custo potencialmente mais baixo. Mas, como uma infraestrutura diferente, não está claro se as nuvens são capazes de executar aplicações científicas com um desempenho razoável por dinheiro gasto.
Proposta: Avaliar a capacidade de uma nuvem em ter um bom desempenho, bem como avaliar o custo da nuvem em termos de desempenho bruto e desempenho de aplicações científicas. Além disso, são avaliados outros serviços, incluindo S3, EBS e DynamoDB, a fim de avaliar as habilidades daqueles a serem utilizados por aplicações e estruturas científicas. Também são avaliadas aplicações de computação científica reais por meio do sistema de *scripts* paralelos em escala Swift.
Validação: Foi verificado o desempenho bruto do EC2 com a execução de micro *benchmarks* para medir o desempenho bruto de diferentes tipos de instância, em comparação com o pico de desempenho teórico reivindicado pelo provedor de recursos. Também se comparou o desempenho real com um sistema não virtualizado típico para entender melhor o efeito da virtualização.
Trabalhos Futuros: Não apresenta trabalhos futuros.

- **Amazon Elastic Compute Cloud (EC2) versus In-House HPC Platform: A Cost Analysis [26]**

  Problema: Embora os centros de computação de alto desempenho (HPC) evoluam continuamente para fornecer mais poder de computação a seus usuários, observa-se um desejo de convergência entre plataformas de computação em nuvem (CC) e computação de alto desempenho (HPC). Excluindo-se o ponto de vista do desempenho, em que muitos estudos destacam uma sobrecarga induzida pela camada de virtualização no núcleo dos *middlewares* em nuvem ao executar uma carga de trabalho HPC, a relação custo-benefício real costuma ser deixada de lado com o desejo de que as instâncias oferecidas pelos provedores de nuvem sejam competitivas do ponto de vista dos custos.
  Proposta: Analisar os elementos que compõe o custo total de propriedade (TCO) de uma instalação interna de HPC, operada desde 2007. A partir do modelo de TCO, comparar os custos necessários para executar a mesma plataforma (e a mesma carga de trabalho) em uma oferta competitiva de nuvem do tipo IaaS. Uma abordagem tripla para comparação de preços é utilizada. Primeiro, é proposto um modelo de preço-desempenho teórico baseado no estudo das instâncias da Amazon EC2. Em seguida, com base na análise de custo total de propriedade da plataforma HPC, é feita uma comparação horária de preços entre o *cluster* interno e as instâncias equivalentes da EC2. Por fim, com base no *benchmarking* experimental no *cluster* local e nas instâncias da nuvem, foi proposta uma atualização do antigo modelo teórico de preços para refletir o desempenho real do sistema.
  Validação: Comparação entre as instâncias EC2 e os nós do *cluster* local. A partir dessa comparação, o modelo de custo é refinado, integrando a pontuação de referência real na equação do modelo. Para esse fim, foi utilizado o *benchmark High Performance Conjugate Gradients* (HPCG.
  Trabalhos Futuros: Estender a análise sobre instâncias do tipo *spot* que permitem fazer lances pelo preço dos recursos. Isso oferece uma chance de uma melhor economia de custos nos preços das taxas de instância e, portanto, pode indicar outras classes de recursos HPC para os quais a opção de aluguel faz sentido. Integrar o custo real de uma nova sala para servidores HPC à análise TCO, a partir do monitoramento dos custos com construção e implementação. Isso também permitirá atualizar o modelo com relação ao custo das tecnologias de ponta para HPC, como, resfriamento direto líquido e interconexões Infiniband.

- **Exploring Instance Heterogeneity in Public Cloud Providers for HPC Applications [27]**

  Problema: A execução de grandes aplicações paralelas (como as de HPC) tornou-se um aspecto importante da computação em nuvem nos últimos anos. Com estas execuções, os usuários podem se beneficiar de custos iniciais mais baixos, maior flexibilidade e atualizações de hardware mais rápidas em comparação com os *clusters* tradicionais. No entanto, o desempenho bruto e a eficiência de custos para uso a longo prazo podem ser uma desvantagem.
  Proposta: Analisar três provedores de nuvem diferentes (Microsoft Azure, Amazon AWS e Google Cloud), por meio da aplicação paralela ImbBench, em termos de adequação a uma execução tão heterogênea de aplicações paralelas grandes. O ImbBench é um aplicativo baseado em MPI que pode criar diferentes padrões de desequilíbrio no uso da CPU e da memória que imita o comportamento de aplicações HPC do mundo real.
  Validação: Execução do *benchmark* ImbBench em infraestruturas compostas por *clusters* com 32 núcleos, formados por quatro instâncias com oito *cores* cada. As instâncias foram criadas nos três provedores escolhidos para este trabalho. Para a métrica de eficiência de custos foi utilizada a metodologia descrita em trabalhos anteriores do autor.
  Trabalhos Futuros: Estender o *benchmark* ImbBench, adicionando suporte para operações de entrada e saída e comunicação de rede a fim de avaliar esses aspectos em termos de heterogeneidade. Também acrescentar suporte para combinar diferentes tipos de operações para melhor representar aplicações do mundo real. Além disso, fornecer uma maneira automática de combinação de instâncias da nuvem para um comportamento de aplicação específico.

3.4 Considerações

Com a conclusão da Revisão Sistemática da Literatura foi possível identificar doze trabalhos que abordam temas relacionados à análise de investimentos em infraestruturas de nuvem. Os trabalhos levam em consideração a execução de aplicações com demandas de processamento de alto desempenho em ambientes de nuvem, bem como a possibilidade de migração destas aplicações, a partir de estruturas clássicas de HPC, para ambientes de nuvem e os custos financeiros envolvidos. Na Seção 4 estes artigos são discutidos como trabalhos relacionados ao tema da pesquisa.

4. Sumarização dos Trabalhos

Esta seção apresenta uma discussão dos trabalhos identificados na literatura, relacionados diretamente às questões que tratam aspectos ligados à adoção de uma solução em nuvem para o aplicações HPC.

No mercado de computação em nuvem, o valor cobrado por cada solução varia muito. Levar em conta somente o desempenho na comparação de fornecedores pode não ser suficiente. Em seu trabalho, [2] (T01, conforme Tabela 8) define uma métrica de eficiência de custos que se propõe a realizar uma comparação mais justa em relação ao que é disponibilizado ao usuário, escalando o valor do desempenho com o preço por hora. Suas conclusões mostram que nuvens podem fornecer uma plataforma viável para a execução de aplicações HPC, mesmo que com algumas desvantagens na implantação, como criação e personalização de instâncias virtuais, problemas de conexão e gerenciamento e tempo de inicialização. Em vários *benchmarks*, os provedores de nuvem obtiveram desempenho e eficiência de custos melhores que o *cluster* local. Além disso, é necessário analisar o comportamento das aplicações de destino, bem como características dos provedores, a fim de escolher o mais adequado para uma determinada aplicação.

Em [18] (T02), são comparadas instâncias da Amazon EC2 com *clusters* locais na execução de um conjunto de *benchmarks*. O índice considerado é o tempo de resposta. Com relação aos tempos de espera em filas, no modelo, foram utilizados traços de execuções reais. Para que fosse possível uma comparação dos custos totais das execuções, foi desenvolvido um modelo econômico com o objetivo de precificar os *clusters*, supondo que estes fossem oferecidos como recursos de nuvem. Por fim, os autores concluem que *clusters* HPC de ponta são superiores em desempenho e que, a partir da perspectiva de usuário, há várias considerações na escolha de uma plataforma, como tempo de espera e custo real.

O trabalho de [19] (T03) apresenta um modelo de desempenho/custo que permite ao usuário decidir qual infraestrutura, local ou em nuvem pública, é ideal para um determinado tipo e tamanho de problema. Quanto maior o número de recursos computacionais ou maior o tempo de execução, maior o custo total, mesmo para estruturas locais não utilizadas. Um exemplo são aplicações que fazem uso intensivo de aceleradores baseados em GPUs. Tal condição pode sobrecarregar o orçamento de uma instituição ao processar grandes quantidades de dados em um ambiente de nuvem ou gerar desperdícios financeiros devido à subutilização em uma infraestrutura local. A principal conclusão é que o uso de máquinas locais, por ano, deve ser bastante alto, algo entre 50% e 100%, para ser rentável. Do contrário, a computação em nuvem é uma alternativa mais econômica que a computação local se o uso de recursos estiver abaixo desses valores.

Shen e seu grupo, [20] (T04), propuseram uma abordagem para provisionamento de recursos, baseada em preços, capaz de atingir uma meta de economia de custos para provedores de serviços de nuvem. A abordagem proposta fornece um conjunto de algoritmos de previsão junto com um modelo auto-regressivo padrão para facilitar a necessidade de previsão de cargas de trabalhos. Para estimar a relação entre carga de trabalho prevista e quantidade de máquinas virtuais, foram realizadas simulações de instâncias da Amazon EC2. Utilizando as políticas de preços da Amazon, foi obtida uma solução de economia de custos para fornecedores de aplicativos. Os resultados do experimento demonstram que esta abordagem é mais econômica em comparação com outras soluções de provisionamento.

Uma avaliação detalhada e abrangente de desempenho e custo de execução de um conjunto de aplicações HPC em uma diversidade de plataformas, variando desde supercomputadores às nuvens computacionais foi apresentada por [10] (T05). Este estudo oportuniza uma visão global que busca responder ao questionamento *por que* e *quem* deve escolher uma nuvem para execução de aplicações HPC e *quais* aplicações e *como* as nuvens devem ser utilizadas para HPC. Também são investigados os aspectos econômicos da execução em nuvem e discutir por que é desafiador ou gratificante para os provedores de nuvem operar negócios para a HPC em comparação com as aplicações em nuvem tradicionais. Deste estudo, algumas lições podem ser observadas: *i*) nuvens podem complementar, com sucesso, supercomputadores, porém substituí-los totalmente ainda é inviável. *Cloud bursting* é uma solução promissora; *ii*) para uma execução de alto desempenho eficiente em nuvem, as aplicações HPC precisam estar cientes do ambiente de nuvem e a nuvem, por sua vez, deve estar preparada para executar aplicações com demandas de alto desempenho; e *iii*) os benefícios econômicos são substanciais, porém, as análises de custo/desempenho para aplicações HPC não são uma tarefa trivial.

Em seu trabalho, Prukkantragorn e Tientanopajai [21] (T06), investigaram a eficiência de valores cobrados que podem beneficiar o cliente de serviços em nuvem no processo de escolha dos pacotes de otimizações oferecidos pelos provedores. Foram estudados os valores praticados pelo provedor de serviços Amazon para a execução de cargas de trabalho de computação de alto desempenho. Ao final, este trabalho apresenta a instância de tamanho adequado para o uso de HPC em diferentes cargas de trabalho e proporções entre o valor cobrado e o tempo de execução reduzido, destacando que a decisão na escolha de um pacote depende da satisfação e uso

dos clientes.

No trabalho de [22] (T07) são apresentados dois algoritmos, o *Proportional Deadline Constrained* (PDC) e o *Deadline Constrained Critical Path* (DCCP) que abordam o problema de escalonamento de fluxos de trabalho nos recursos de nuvem provisionados dinamicamente. Em termos de desempenho de custo, em geral, os algoritmos PDC e DCCP retornaram o menor custo de computação, em todos os fluxos de trabalho e configurações de instância. No geral, ambos os algoritmos são capazes de obter altas taxas de sucesso, enquanto na maioria dos casos apresentam o menor custo geral por uso.

Os últimos anos conduziram a um rápido aumento no número de tipos de configurações de hardware de computação em nuvens públicas e opções de valores oferecidos aos usuários, conforme nos demonstra [23] (T08) em seu estudo. Tal movimentação dificulta a análise de qual opção se torna mais vantajosa, economicamente, em uma migração de aplicações de uso geral para a nuvem. Com isso, o autor investiga se esta mesma análise pode ser estendida para execuções de HPC. Com o uso de uma configuração de hardware clássica para HPC, foi realizada uma comparação do custo total das operações de vários provedores de nuvem pública e privada de HPC. A análise mostrou sob quais condições operacionais a opção de nuvem pública pode ser uma alternativa mais econômica para aplicações do tipo HPC.

Em [24] (T09), os autores buscaram identificar as instâncias de máquina virtual, em nuvens públicas disponíveis, que possam ser adequadas para aplicações HPC, tanto em termos de desempenho quanto de eficiência de custos. Também foram avaliados quais tipos de aplicações podem se beneficiar da execução na nuvem. Para isso, é preciso analisar as características das instâncias, levando em consideração os aspectos relevantes para HPC. Também é necessária uma análise da eficiência de custos usando os *benchmarks* tradicionais de HPC. Os resultados mostraram que o desempenho de rede continua sendo um gargalo significativo para o desempenho das aplicações. Além disso, pagar por uma nuvem mais poderosa nem sempre garante melhorias e pode até levar a reduções de desempenho. Isso deve ao fato de que as aplicações HPC podem ter características de escalonamento não suportadas pelo ambiente de nuvem, além de sofrer com os efeitos de limitações próprias de ambientes virtualizados, comumente utilizados em nuvens computacionais.

Com foco em aplicações científicas, os autores de [25] (T10) realizaram uma avaliação das instâncias da Amazon EC2, com o objetivo de executar aplicações com desempenho satisfatório e com custo potencialmente mais baixo. Em comparação das instâncias de nuvem pública com as de uma nuvem privada, foi constatado que a eficiência e o desempenho das duas estruturas eram bastante similares. Quanto aos custos, as instâncias virtuais otimizadas para computação são as que apresentaram melhor eficiência financeira. O maior gargalo detectado foi com relação à comunicação de rede que pode afetar, diretamente, a execução satisfatória de aplicações HPC.

No seu trabalho, [26] (T11) propôs analisar os elementos que compõe o custo total de propriedade (TCO) de uma estrutura interna de HPC, operando desde 2007 e, de posse das informações, comparar os custos necessários para executar a mesma carga de trabalho em uma estrutura de nuvem pública. Os resultados obtidos mostram que a migração de cargas de trabalho HPC para nuvem não é apenas um problema de adaptabilidade do desempenho da nuvem às necessidades das aplicações HPC, mas também um problema ao determinar corretamente quais tipos de trabalho são bons candidatos a serem executados na nuvem para evitar sobrecarga de custos.

Em seu estudo, [27] (T12) analisou três provedores de nuvens públicas: Microsoft Azure, Amazon AWS e Google Cloud. Com uso de uma aplicação paralela ImbBench, avaliou a adequação de uma execução heterogênea de aplicações paralelas grandes. A avaliação de eficiência de custo foi realizada por meio da metodologia apresentada em [2]. Os resultados destacam que a execução heterogênea é mais benéfica na plataforma Azure, com uma eficiência de custos de até 50% em comparação com a execução em instâncias homogêneas, mantendo o mesmo desempenho. Os outros dois provedores são menos adequados, pois o tipo de instância mais barato também é o mais rápido, para o caso da Amazon AWS, ou o provedor oferece apenas instâncias que variam no tamanho da memória, mas não no desempenho, como é o caso do provedor Google Cloud.

A Tabela 9 sumariza os trabalhos relacionados nesta seção. São apresentadas características identificadas nos trabalhos para um melhor entendimento do posicionamento de cada um deles em relação aos demais. A primeira coluna desta tabela identifica os trabalhos que tratam diretamente de assuntos relacionados às execuções HPC. Nas colunas "Simulação" e "Infra. Real" são marcados os trabalhos de acordo com a técnica utilizada para validação das propostas. Na sequência, as colunas "Traço" e "*Benchmark*" indicam os tipos de origem dos dados usados para a validações.

Nas colunas seguintes, "Nuvem Privada", "Nuvem Pública" e "*Cluster* Local" são referenciados os ambientes de testes utilizados para execução dos experimentos apresentados nos trabalhos. A coluna "Análise de Custo" indica os trabalhos que apresentam análises de custo que considerem aplicações HPC suas execuções em ambientes de nuvem. Por fim, as colunas "Satisf. Usuário" e "Satisf. Provedor" caracteriza se o trabalho considera a satisfação do usuário, do provedor de nuvem ou de ambos.

5. Síntese

Nesta seção, o estudo decorrente da leitura dos trabalhos selecionados é sintetizado de forma a responder as questões de pesquisa que motivaram a realização da presente RSL. Estas questões são, portanto, retomadas e respondidas tendo como base o conhecimento absorvido.

QP 1: Qual o impacto econômico nas decisões de implantação de aplicações de alto desempenho em nuvens

Tabela 9. Abordagens adotadas pelos trabalhos relacionados

T#	HPC	Simulação	Infra. Real	Traço	*Benchmark*	Nuvem Privada	Nuvem Pública	*Cluster* Local	Análise de Custo	Satisf. Usuário	Satisf. Provedor
T01	✓		✓		✓		✓	✓	✓	✓	
T02	✓		✓	✓	✓		✓	✓	✓	✓	
T03	✓		✓				✓	✓	✓	✓	
T04		✓				✓	✓		✓	✓	✓
T05	✓	✓	✓		✓	✓	✓	✓		✓	✓
T06	✓		✓		✓		✓		✓	✓	
T07		✓		✓						✓	
T08	✓					✓	✓	✓	✓	✓	
T09	✓		✓		✓		✓	✓		✓	
T10	✓		✓		✓	✓	✓		✓	✓	
T11	✓				✓		✓	✓	✓		
T12	✓		✓		✓		✓			✓	

computacionais?

Esta QP tem o objetivo de identificar, considerando custos financeiros, como instituições podem se beneficiar da adoção de nuvens. Os trabalhos [2, 25] exemplificam estes ganhos caracterizando as aplicações e investigando aspectos econômicos das execuções HPC em ambientes de nuvem. Também são elencados argumentos para determinar quando aplicações de alto desempenho podem, realmente, obter vantagens de nuvens computacionais em detrimento de infraestruturas locais dedicadas.

QP 2: Como os usuários podem identificar qual a melhor configuração de nuvem, seja pública ou privada, para executar suas aplicações?

A migração para um ambiente de nuvem computacional requer, da parte de instituição, investimentos na nova infraestrutura. Como caracterizado em [23, 19], a análise de custos deve tanto identificar o tipo de nuvem a ser implantado, privada ou pública, não desconsiderando a hipótese de uma implementação híbrida [10, 25]. Em alguns casos, a análise dos custos utiliza dados de desempenho obtidos por simulação [20, 10, 22] ou por experimentos envolvendo o uso de uma infraestrutura existente, explorando a execução de benchmarks [2, 18, 10, 21, 24, 25, 26, 27] ou analisando o comportamento da execução [18, 22]. Este tipo de estudo requer grande envolvimento de pessoal, seja na elaboração do processo de simulação ou de coleta de dados de execuções, seja na análise e interpretação dos resultados. A literatura também apresenta, como em [2, 18, 19, 20, 21, 23, 25, 26], modelos de custos analíticos, cujo esforço de aplicação é, quando comparado aos anteriormente citados, menor. O aspecto relevante a ser considerado, neste caso, é identificar o grau de precisão do modelo a ser utilizado.

QP 3: Quais modelos de custo financeiro são empregados na execução de aplicações de alto desempenho em nuvens?

Os trabalhos identificados na RSL empregam informações sobre o desempenho das aplicações na análise do impacto do investimento realizado. Nestes trabalhos [2, 18, 10, 24, 25, 21, 26], a métrica sobre desempenho pode ser entendida como principal componente do modelo de custo. Destaca-se que, para análise do investimento e do seu impacto no desempenho da aplicação, o modelo de análise de custo TCO foi o mais utilizado. Também foi identificado que, à exceção do trabalho T11 [26], os modelos de custo apresentados estão voltados para responder às necessidades dos usuários sobre a análise dos custos. Os trabalhos [20] e [10] apresentaram uma análise de custo sobre a ótica do provedor.

QP 4: Como identificar eventuais perdas financeiras em função de imprecisão no dimensionamento de infraestruturas de nuvens para aplicações com demandas de HPC?

Nos trabalhos identificados na RSL, os casos de estudo relatados nos trabalhos selecionados apresentam a instanciação de uma aplicação na nuvem e avaliação de seu comportamento. A perda financeira é analisada pela avaliação da estimativa do desempenho das aplicações sobre um conjunto de recursos alocados. O modelo de decisão baseado em custos, TCO, é o mais utilizado para apoiar esta análise.

6. Conclusão e Oportunidades de Pesquisa

A consolidação das tecnologias de computação em nuvem promoveu um grande crescimento no interesse por ambientes capazes de suportar a execução de aplicações que necessitam de alto desempenho (HPC). Conforme apresentado neste trabalho, percebe-se que a migração das aplicações, a partir de estruturas locais dedicadas para nuvens, não é uma tarefa fácil e traz alguns desafios, principalmente no que diz respeito aos custos financeiros envolvidos no processo. Neste estudo, uma revisão sistemática da literatura buscou temas relaciona-

dos à análise de investimentos em infraestruturas de nuvens computacionais e quais aspectos devem ser levados em consideração frente aos novos desafios impostos pela migração de aplicações HPC para ambientes de nuvem.

Com a conclusão da revisão sistemática da literatura foi possível identificar doze trabalhos que abordam temas relacionados à análise de investimentos em infraestruturas de nuvem. Os trabalhos levam em consideração a execução de aplicações com demandas de processamento de alto desempenho em ambientes de nuvem, bem como a possibilidade de migração destas aplicações, a partir de estruturas clássicas de HPC, para ambientes de nuvem, juntamente comos custos financeiros envolvidos.

No entanto, na adoção de uma infraestrutura de nuvem, eventuais perdas financeiras não são resultado apenas do sub ou superdimensionamento dos recursos alocados. Outros aspectos podem ser relevantes no contexto, como a questão de privacidade das informações [28, 29] e também o suporte à aplicações de missão crítica [30] que limitam o horizonte de opções de implantação da infraestrutura de suporte.

Especificamente sobre custos relacionados à infraestrutura, observa-se que não emergiram considerações sobre os custos de comunicação associados às transferências de dados nem à adoção de soluções de nuvens híbridas. Estes aspectos se mostram como oportunidades de pesquisa em aberto. Outra consideração a ser feita é que os trabalhos selecionados, embora focados em aplicações HPC, não consideram características específicas à determinadas categorias de instituições, como industrial, comercial, acadêmica ou de pesquisa, na análise dos investimentos. Entende-se que a natureza das instituições deva impactar na análise dos resultados financeiros, pois é possível que, como pode ser o caso em instituições acadêmicas e de pesquisa, a análise em termos do resultado financeiro imediato pode não ser suficiente. Um estudo aprofundado sobre o uso de infraestruturas de nuvem em ambientes acadêmicos e de pesquisa se apresenta, assim, como um tema a ser investigado.

Reconhecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Contribuições dos autores

- Maicon Ança dos Santos: escrita do texto e revisão sistemática da literatura.
- Gerson Geraldo H. Cavalheiro: orientador da pesquisa.

Referências

[1] MELL, P.; GRANCE, T. *The NIST definition of cloud computing*. Gaithersburg, MD, 2011. 7 p. Disponível em: <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf>.

[2] ROLOFF, E. et al. High Performance Computing in the cloud: Deployment, performance and cost efficiency. In: *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*. Taipei, Taiwan: IEEE, 2012. p. 371–378.

[3] ALADYSHEV, O. S. et al. Variants of deployment the high performance computing in clouds. In: *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. Moscow: IEEE, 2018. p. 1453–1457. Disponível em: <http://ieeexplore.ieee.org/document/8317371/>.

[4] NETTO, M. A. S. et al. HPC Cloud for Scientific and Business Applications: Taxonomy, Vision, and Research Challenges. *ACM Computing Surveys*, v. 51, n. 1, p. 1–29, abr. 2018. ArXiv: 1710.08731. Disponível em: <http://arxiv.org/abs/1710.08731>.

[5] AL-ROOMI, M. et al. Cloud Computing Pricing Models: A Survey. *International Journal of Grid and Distributed Computing*, v. 6, n. 5, p. 93–106, out. 2013.

[6] STREBEL, J.; STAGE, A. An economic decision model for business software application deployment on hybrid Cloud environments. *IT Performance Management*, p. 13, 2010.

[7] MIERITZ, L.; KIRWIN, B. Defining Gartner Total Cost of Ownership. p. 11, 2005.

[8] FILIOPOULOU, E. et al. Integrating cost analysis in the cloud: A SoS approach. In: *2015 11th International Conference on Innovations in Information Technology (IIT)*. Dubai, United Arab Emirates: IEEE, 2015. p. 278–283.

[9] WALTERBUSCH, M.; MARTENS, B.; TEUTEBERG, F. Evaluating cloud computing services from a total cost of ownership perspective. *Management Research Review*, v. 36, n. 6, p. 613–638, maio 2013.

[10] GUPTA, A. et al. Evaluating and Improving the Performance and Scheduling of HPC Applications in Cloud. *IEEE Transactions on Cloud Computing*, v. 4, n. 3, p. 307–321, jul. 2016.

[11] MANSOURI, Y.; PROKHORENKO, V.; BABAR, M. A. An Automated Implementation of Hybrid Cloud for Performance Evaluation of Distributed Databases. *arXiv:2006.02833 [cs]*, jun. 2020. ArXiv: 2006.02833. Disponível em: <http://arxiv.org/abs/2006.02833>.

[12] YELICK, K. et al. The Magellan Report on Cloud Computing for Science. *U.S. Department of Energy - Office of Science - Office of Advanced Scientific Computing Research (ASCR)*, p. 170, dez. 2011. Disponível em: <http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Magellan_Final_Report.pdf>.

[13] PARASHAR, M. et al. Cloud Paradigms and Practices for Computational and Data-Enabled Science and Engineering. *Computing in Science & Engineering*, v. 15, n. 4, p. 10–18, jul.

2013. Disponível em: <http://ieeexplore.ieee.org/document/6530588/>.

[14] GANTIKOW, H. et al. A Taxonomy for HPC-aware Cloud Computing. In: . [S.l.: s.n.], 2015. (2nd Baden-Württemberg Center of Applied Research Symposium on Information and Communication Systems SInCom 2015, Konstanz), p. 57 – 62.

[15] ZHAO, H.; LI, X. Designing Flexible Resource Rental Models for Implementing HPC-as-a-Service in Cloud. In: *2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum*. Shanghai, China: IEEE, 2012. p. 2550–2553.

[16] XIAO, Y.; WATSON, M. Guidance on Conducting a Systematic Literature Review. *Journal of Planning Education and Research*, v. 39, n. 1, p. 93–112, mar. 2017.

[17] OKOLI, C.; SCHABRAM, K. A Guide to Conducting a Systematic Literature Review of Information Systems Research. *SSRN Electronic Journal*, 2010.

[18] MARATHE, A. et al. A comparative study of high-performance computing on the cloud. In: *Proceedings of the 22nd International Symposium on High-Performance Parallel and Distributed Computing - HPDC '13*. New York, New York, USA: ACM Press, 2013. p. 239.

[19] GUERRERO, G. D. et al. A Performance/Cost Model for a CUDA Drug Discovery Application on Physical and Public Cloud Infrastructures. *Concurrency and Computation: Practice and Experience*, v. 26, n. 10, p. 1787–1798, jul. 2014.

[20] SHEN, Y. et al. Cost-Optimized Resource Provision for Cloud Applications. In: *2014 IEEE Intl Conf on High Performance Computing and Communications, 2014 IEEE 6th Intl Symp on Cyberspace Safety and Security, 2014 IEEE 11th Intl Conf on Embedded Software and Syst (HPCC,CSS,ICESS)*. Paris, France: IEEE, 2014. p. 1060–1067.

[21] PRUKKANTRAGORN, P.; TIENTANOPAJAI, K. Price efficiency in High Performance Computing on Amazon Elastic Compute Cloud provider in Compute Optimize packages. In: *2016 International Computer Science and Engineering Conference (ICSEC)*. Chiang Mai, Thailand: IEEE, 2016. p. 1–6.

[22] ARABNEJAD, V.; BUBENDORFER, K.; NG, B. Scheduling deadline constrained scientific workflows on dynamically provisioned cloud resources. *Future Generation Computer Systems*, v. 75, p. 348–364, out. 2017.

[23] DREHER, P. et al. Cost Analysis Comparing HPC Public Versus Private Cloud Computing. In: HELFERT, M. et al. (Ed.). *Cloud Computing and Services Science*. Cham: Springer International Publishing, 2017. v. 740, p. 294–316.

[24] ROLOFF, E. et al. HPC Application Performance and Cost Efficiency in the Cloud. In: *2017 25th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*. St. Petersburg, Russia: IEEE, 2017. p. 473–477.

[25] SADOOGHI, I. et al. Understanding the Performance and Potential of Cloud Computing for Scientific Applications. *IEEE Transactions on Cloud Computing*, v. 5, n. 2, p. 358–371, abr. 2017.

[26] EMERAS, J. et al. Amazon Elastic Compute Cloud (EC2) versus In-House HPC Platform: A Cost Analysis. *IEEE Transactions on Cloud Computing*, v. 7, n. 2, p. 456–468, abr. 2019.

[27] ROLOFF, E. et al. Exploring Instance Heterogeneity in Public Cloud Providers for HPC Applications:. In: *Proceedings of the 9th International Conference on Cloud Computing and Services Science*. Heraklion, Crete, Greece: SCITEPRESS - Science and Technology Publications, 2019. p. 210–222.

[28] RAMGOVIND, S.; ELOFF, M. M.; SMITH, E. The management of security in Cloud computing. In: *2010 Information Security for South Africa*. Johannesburg, South Africa: IEEE, 2010. p. 1–7. Disponível em: <http://ieeexplore.ieee.org/document/5588290/>.

[29] BHAVANI, P.; JYOTHI, C. Investigation on security challenges over a cloud computing. *International Journal of Scientific Research in Science and Technology*, v. 3, p. 1374–1380, 2017.

[30] FICCO, M.; AMATO, A.; VENTICINQUE, S. Hosting Mission-Critical Applications on Cloud: Technical Issues and Challenges. In: LAMBOGLIA, R. et al. (Ed.). *Network, Smart and Open*. Cham: Springer International Publishing, 2018. v. 24, p. 179–191. Series Title: Lecture Notes in Information Systems and Organisation. Disponível em: <http://link.springer.com/10.1007/978-3-319-62636-9_12>.