UNIVERSIDADE FEDERAL DO CEARÁ
CURSO DE GRADUAÇÃO EM REDES DE COMPUTADORES

CARLOS BRUNO PEREIRA BEZERRA

PREDIÇÃO DE CARGAS PARA PROVISIONAMENTO DE RECURSOS EM COMPUTAÇÃO EM NUVEM

QUIXADÁ - CEARÁ

2013

CARLOS BRUNO PEREIRA BEZERRA

PREDIÇÃO DE CARGAS PARA PROVISIONAMENTO DE RECURSOS EM COMPUTAÇÃO EM NUVEM

Monografia apresentada no Curso de Graduação em Redes de Computadores da Universidade Federal do Ceará, como requisito parcial para obtenção do grau de Tecnólogo em Redes de Computadores.

Orientador: Jeandro de Mesquita Bezerra

QUIXADÁ - CEARÁ

2013

C824p Bezerra, Carlos Bruno.

Predição de cargas para Provisionamento de Recursos em Computação em Nuvem / Carlos Bruno Pereira Bezerra. – Quixadá, 2013.

49 p.;il.

Orientador: Prof. Msc. Jeandro de Mesquita Bezerra

Monografia (Graduação em Redes de Computadores) - Universidade Federal do Ceará, Centro de Ciências e Tecnologia.

1. Computação em Nuvem 2. Computação Verde 3. Tecnologia da Informação 4. Virtualização 5.

I. Universidade Federal do Ceará, Centro de Ciências e Tecnologia.

CDD:001.6

CARLOS BRUNO PEREIRA BEZERRA

PREDIÇÃO DE CARGAS PARA PROVISIONAMENTO DE RECURSOS EM COMPUTAÇÃO EM NUVEM

Monografia apresentada no Curso de Graduação em Redes de Computadores da Universidade Federal do Ceará, como requisito parcial para obtenção do grau de Tecnólogo.

Aprovada em: 01/08/2013

BANCA EXAMINADORA

Prof. Msc. Jeandro de Mesquita Bezerra
Universidade Federal do Ceará – UFC
Orientador

Prof. Msc. Paulo Antônio Leal Rego
Universidade Federal do Ceará – UFC

Prof. Dr. Flávio R. C. Sousa
Universidade Federal do Ceará – UFC

AGRADECIMENTOS

À minha mãe, pelo incentivo e apoio nos momentos difíceis.

A todos os meus amigos, em especial aos taberneiros Jard MacLeod, Rocha Batera, Krisninha, Marquim Baixista e o não menos nórdico, Juin Bezerra.

Ao Maiden, por me proporcionar momentos de reflexões enquanto escrevia este trabalho.

Ao professor Jeandro Mesquita pela orientação e oportunidade de iniciação à pesquisa.

“Little by little, one travels far.”
J.R.R. Tolkien

RESUMO

Desde a eclosão da internet nos anos 90, a área da computação vem sofrendo grandes mudanças, todas acompanhadas com as necessidades de uma melhor experiência dos seus usuários. Nesse contexto, a humanidade tomou conhecimento de novas possibilidades para ampliar seus negócios. Sendo assim, a Computação em Nuvem adveio com o propósito de oferecer recursos computacionais como serviços, reduzindo os gastos com aquisição de produtos relacionados à Tecnologia da Informação. Computação em Nuvem é um modelo de computação atraente que permite que os recursos sejam providos de acordo com a demanda, ou seja, os usuários podem locar os recursos da nuvem que forem necessários. Ainda que a Computação em Nuvem traga muitos benefícios aos seus usuários, a construção e manutenção de grandes *data centers* mostram-se caros. Dentre as principais preocupações acerca das despesas, encontra-se o consumo elevado de energia elétrica, que além dos gastos, também agride ao meio ambiente. Para amenizar esses problemas, surgiu-se o modelo chamado Computação Verde, que tem como um dos seus objetivos a proposta de implantação de *data centers* que consumam menos energia e gerem menos poluição. Tendo como base este contexto, este trabalho propõe um algoritmo que facilitará as tomadas de decisões dos sistemas de gerenciamento de recursos dos *data centers*, no que diz respeito ao desligamento de servidores subutilizados e as demais decisões que tenham como objetivo atender à Computação Verde.

Palavras-Chave: Computação em Nuvem. Computação Verde. Tecnologia da Informação. Virtualização.

.

ABSTRACT

Since the outbreak of the Internet in the 90s, the area of computing has experimented great changes, accompanied with all the needs for a better experience of its users. In this context, humanity has became aware of new possibilities to expand their business. Thus, Cloud Computing came with the purpose of providing computing resources as services, cheapening the cost of purchase of products related to Information Technology. Cloud computing is an attractive computing model that allows resources to be provided according to the demand, i.e., users can rent the resources from the cloud as necessary. Though cloud computing brings many benefits to its users, the construction and maintenance of large data centers show up expensive. The high consumption of electricity is among the main concerns about the costs, that besides spending also damages the environment. To alleviate these problems, came up a model called Green Computing, which has as one of its objectives the proposed deployment of data centers that consume less power and generate less pollution. Based on this context, this paper proposes an algorithm that will facilitate the management systems decisions in data centers, concerning to the shutdown of underutilized servers and other decisions that aim to meet the Green Computing.

Keywords: Cloud Computing. Green Computing. Information Technology. Virtualization.

.

LISTA DE FIGURAS

LISTA DE SIGLAS

TI	Tecnologia da Informação
SLA	*Service Level Agreement*
API	*Application Programming Interface*
MV	Máquinas Virtual
NIST	*National Institute of Standards and Technology*
MF	Máquina Física
VMM	*Virtual Machine Monitor*
XML	*Extensible Markup Language*
SOAP	*Simple Object Access Protocol*
CaaS	Comunicação como Serviço
MaaS	Gerenciamento como Serviço
DaaS	Dados como Serviço
ITaaS	TI como Serviço
XaaS	Tudo como Serviço
IaaS	Infraestrutura como Serviço
PaaS	Plataforma como Serviço
SaaS	*Software* como Serviço
CPU	*Central Processing Unit*
RAM	*Random Access Memory*
P2P	*Peer-to-Peer*
DDoS	*Distributed Denial of Service*
DHCP	*Dynamic Host Configuration Protocol*
QoS	*Quality of Service*

SUMÁRIO

1 INTRODUÇÃO

Computação em Nuvem ou *Cloud Computing* é um novo paradigma na área da computação que proporciona serviços de TI sob demanda. Os usuários têm a vantagem de usufruir dos recursos oferecidos pelos provedores de nuvem pagando apenas por aquilo que usar, tendo um acordo de nível de serviço ou SLA garantido pelo provedor. O não cumprimento do acordo, como indisponibilidade de algum recurso computacional, pode acarretar em penalidades.

A Computação em Nuvem atende tanto às empresas que pretendem migrar toda sua infraestrutura tecnológica para um provedor de nuvem, como também usuários que almejam utilizar serviços da nuvem para armazenar seus arquivos e desenvolver aplicações utilizando toda a plataforma e API remotamente.

Apesar dos inúmeros benefícios, os *data centers* tradicionais de *Cloud Computing* necessitam de sistemas de gerência que visem a redução dos custos operacionais, como por exemplo, a atenuação do consumo energético com o desligamento ou migração de MV subutilizadas. Isso pode ser alcançado com melhorias significativas nos sistemas que gerenciam a nuvem computacional.

1.1 Motivação

A principal motivação deste trabalho é contribuir com a melhoria dos sistemas de gerenciamento presentes nos *data centers* tradicionais de Computação em Nuvem. Especificamente, ajudar com a redução dos custos operacionais.

1.2 Obejetivos

Essa seção visa descrever os objetivos gerais e específicos que devem ser atendidos por este trabalho.

1.2.1 Objetivo Geral

O presente trabalho tem como objetivo realizar predições das cargas de trabalho[1] em *data centers*. A proposta pode auxiliar na distribuição das cargas e facilitar as tomadas de decisões dos sistemas de gerenciamento no paradigma de Computação em Nuvem.

[1] Conjunto de *jobs* e *tasks* (ou tarefas) que é processado em um ambiente computacional. Cada task requer uma quantidade mínima de recurso (CPU, memória) para ser processada (MISHRA et al., 2009).

1.2.2 Objetivos específicos

1. Implementar um algoritmo para a predição das cargas de trabalho;
2. Facilitar as tomadas de decisões dos sistemas de gerenciamento;
3. Contribuir com a redução do consumo energético em *data centers*;
4. Propor uma funcionalidade adicional na nuvem, como apoio aos sistemas de gerência;

1.3 Organização da monografia

Após a introdução, o trabalho está dividido em mais quatro capítulos. O capítulo 2 levanta os principais conceitos no que concerne o novo paradigma de computação distribuída denominado Computação em Nuvem e compõe o embasamento teórico deste trabalho. No Capítulo 3 será fundamentada toda a metodologia utilizada, ou seja, a implementação de um algoritmo de predição de cargas de trabalho para os sistemas de gerenciamento de Computação em Nuvem. O Capítulo 4 analisa os resultados obtidos. Enquanto o Capítulo 5 apresenta as considerações finais a respeito do trabalho realizado.

2 COMPUTAÇÃO EM NUVEM

Nesse capítulo, serão abordados aspectos que definem o novo foco de pesquisas e negócios hoje intitulado Computação em Nuvem. As Seções 2.1 e 2.2 trazem as principais definições e apresentam os elementos que compõem a arquitetura deste modelo. Em seguida, são abordados os principais modelos de serviço e os modos de implantação existentes nas Seções 2.3 e 2.4. A seção 2.5 aborda definições sobre a infraestrutura dos *data centers* convencionais, enfatizando seus desafios. As seções 2.6 e 2.7 falam, respectivamente, da Computação Verde e do gerenciamento de recursos no contexto da Computação em Nuvem.

2.1 O que é Computação em Nuvem?

A sociedade humana moderna segue um ritmo acelerado de desenvolvimento, no qual serviços são entregues de uma forma completamente transparente. As atuais infraestruturas de utilidade pública permitem entregar serviços em qualquer lugar e a qualquer hora, baseadas no modelo de pagamento vinculado ao uso, de forma que se pode acender a luz, abrir a torneira ou fazer uma ligação. Desta forma também é como funciona a Computação em Nuvem, pois da mesma maneira os seus usuários pagam pelo serviço de computação utilizado.

A nuvem pode ser entendida como uma abstração de toda a complexidade da infraestrutura subjacente, referindo-se de modo geral a uma combinação de tecnologias e paradigmas (virtualização, Computação Utilitária, Computação em Grade, arquitetura orientada a serviços, etc) , como pode ser visto na Figura 1. Para a utilização dos recursos oferecidos na nuvem, como armazenamento, banco de dados, processamento, entre outros, usuários necessitam dispor apenas de um equipamento computacional (notebook, smartphone, computador pessoal) com acesso à internet, um sistema operacional e um navegador *web* (SOUSA et al., 2009).

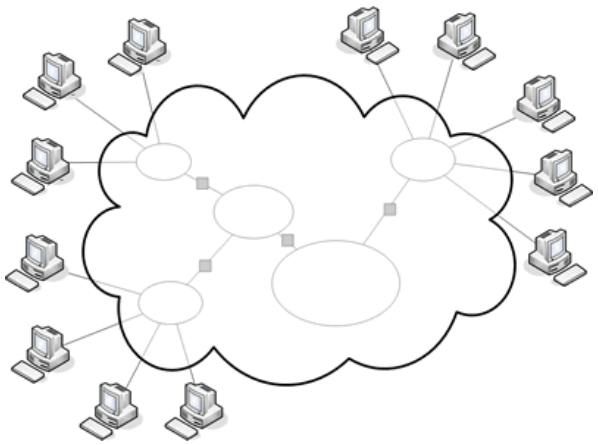

Figura 1: Diagrama de Rede (COULOURIS et al., 2005)

Atualmente a área de Computação em Nuvem tem sido um assunto amplamente dis-

cutido pelo mercado e pela academia. Por ser considerado um novo paradigma, não existem definições claras dos padrões a serem utilizados. Existem, entanto, muitas definições consistentes para as tecnologias que tendem a compor a infraestrutura de nuvem (VERDI et al., 2010).

Segundo o NIST[1], Computação em Nuvem é:

> Um modelo que possibilita acesso, de modo conveniente e sob demanda, a um conjunto de recursos computacionais configuráveis (por exemplo, redes, servidores, armazenamento, aplicações e serviços) que podem ser rapidamente adquiridos e liberados com mínimo esforço gerencial ou interação com o provedor de serviços(BADGER et al., 2011, p. 2-1).

O NIST ainda descreve que o modelo de Computação em Nuvem deve ser composto por cinco características essenciais, três modelos de serviço e quatro modelos de implantação. Entretanto, os modelos de serviço e os modelos de implantação serão introduzidos nas seções 2.4 e 2.5. As cinco características essenciais são descritas a seguir:

Atendimento self-service e sob demanda: O usuário pode alocar recursos computacionais unilateralmente sem a necessidade de interação humana com o provedor do serviço. Empresas que atuam na área de IaaS oferecem uma Application Programming Interface (Interface de Programação de Aplicativos) (API) própria, que pode ser utilizada para a requisição dinâmica de recursos através de scripts personalizados. A Computação em Nuvem não é mais que um modelo de prestação de serviços. Neste tipo de computação, tudo o que pode oferecer um sistema de computação é fornecido como um serviço.

Amplo acesso à rede: Infraestrutura computacional, plataformas de desenvolvimento e aplicações são acessadas via rede através de protocolos padrão. Isto possibilita a utilização dos serviços por máquinas clientes que variam de desktops robustos a dispositivos móveis com severas limitações de recursos. A Computação em Nuvem desenvolve a idéia da internet com aplicações remotas.

Pooling de recursos: O provedor detém um conjunto de recursos físicos e virtuais que são alocados para usuários e liberados pelos mesmos de forma dinâmica e de acordo com a demanda existente. A informação e aplicações já não são dependentes de nossos computadores ou sistema operacional. A alocação de um mesmo recurso é muitas vezes feita para mais de um usuário simultaneamente, prática comumente definida como multi-arrendamento (em inglês, multi-tenancy). Ademais, o usuário não possui controle ou conhecimento preciso do ponto geográfico de origem do serviço que está utilizando.

Elasticidade: Os recursos providos por um ambiente computacional em nuvem são inerentemente escaláveis. O termo elasticidade é utilizado para transmitir a idéia de que o usuário

[1] Agência Governamental da administração de tecnologia do Departamento de Comércio dos Estados Unidos. A missão do instituto é promover a inovação e a competitividade industrial dos Estados Unidos.

pode, a qualquer momento, aumentar ou diminuir a quantidade de recursos utilizados. Essa característica cria a ilusão de que os recursos oferecidos são ilimitados e que o usuário pode fazer uso da quantidade que lhe for conveniente. O consumidor do serviço pode requisitar dinamicamente mais ou menos recursos para sua aplicação para se adaptar à demanda dos seus usuários. Por exemplo, em períodos de pico o consumidor solicita à nuvem mais recursos computacionais (como, por exemplo, servidores adicionais), podendo depois liberar tais recursos, quando os mesmos não forem mais necessários.

Medição de serviços: Os recursos oferecidos podem ser monitorados, controlados e reportados ao provedor e ao usuário de forma transparente. Este mecanismo permite que sejam cobrados valores referentes ao grau de utilização dos recursos e estipulados por meio de contrato entre provedor e usuário. Os consumidores pagam aos provedores de serviço de nuvem de acordo com o consumo efetuado (modelo de pagamento pelo uso semelhante a utilidades como energia e gás).

O modelo de Computação em Nuvem foi desenvolvido com o objetivo de fornecer serviços de fácil acesso, baixo custo e com garantias de disponibilidade e escalabilidade. Este modelo visa fornecer, basicamente, três benefícios. O primeiro benefício é reduzir o custo na aquisição e composição de toda infraestrutura requerida para atender as necessidades das empresas, podendo essa infraestrutura ser composta sob demanda e com recursos heterogêneos e de menor custo. O segundo é a flexibilidade que esse modelo oferece no que diz respeito à adição e substituição de recursos computacionais, podendo escalar tanto em nível de recursos de hardware quanto *software* para atender às necessidades das empresas e usuários. O último benefício é prover uma abstração e facilidade de acesso aos usuários destes serviços. Neste sentido, os usuários dos serviços não precisam conhecer aspectos de localização física e de entrega dos resultados destes serviços (SOUSA et al., 2009).

Esta mudança para um modelo centrado no servidor não traz vantagens apenas para o usuário que fica liberado de toda a gerência local necessária para manter as aplicações (intensas configurações e grandes quantidades de backups), mas também traz vantagens para os produtores de *software* (BARROSO; HöLZLE, 2007). O desenvolvimento de aplicações é mais simples pois as mudanças e as melhorias nos *softwares* são feitas em um único local, ao invés de serem feitas em milhões de clientes que possuem diferentes configurações de hardware. Além disso, uma vez que o uso do *software* hospedado no provedor passa a ser compartilhado por diversos usuários, o custo também se divide entre estes usuários, fazendo com que os valores pela utilização dos serviços sejam reduzidos (VERDI et al., 2010).

Atualmente, o mercado da Computação em Nuvem se encontra incrivelmente valorizado, estima-se "superior a 42.000 milhões de dólares em 2012"(MAHOWALD, 2011). Quando fala-se do mercado brasileiro, a Computação em Nuvem vai crescer exponencialmente. O Brasil é o país que mais tem interesse por Computação em Nuvem quando comparado com os demais

da América Latina. Até hoje, 18% das médias e grandes empresas brasileiras já utilizam alguma aplicação de Computação em Nuvem. Estima-se que até 2013, esta fatia deva saltar para 30% a 35%, número que é aproximadamente 60% maior do que a base atual. Nos Estados Unidos, entre 45% e 55% das companhias médias e grandes já utilizam algum serviço de Computação em Nuvem. Na Europa o número está entre 35% e 40% (MAHOWALD, 2011).

Como foi visto anteriormente, a Computação em Nuvem proporciona facilidades, além de baratear o desenvolvimento de *software*, aquisição de computadores e equipamentos de rede, a seus usuários, de modo que eles sempre poderão alocar e liberar recursos da nuvem sem nenhum esforço e exigir uma quantidade mínima de recurso disponível. Para tornar isso possível, a Computação em Nuvem conta com uma série de tecnologias e modelos computacionais.

Na próxima seção, serão abordadas algumas das principais tecnologias que fazem com que o paradigma de *Cloud Computing* entregue serviços de computação de maneira aparentemente descomplicada a seus clientes.

2.2 Arquitetura

Pôde-se notar na seção 2.1 a importância da adoção da Computação em Nuvem nos dias atuais. Não somente devido ao fato de diminuir os gastos com infraestrutura de TI como também facilitar o desenvolvimento de trabalhos relacionados à Tecnologia da Informação. Embora não trivial, os serviços são oferecidos de forma transparente ao usuário. No entanto, a nuvem é somente a forma mais simples de se enxergar a complexidade e heterogeneidade das tecnologias e modelos que compõem a mesma. Vale ressaltar, que o que será discutido nesta seção, não implica em um padrão de tecnologias e modelos, tendo em mente que este é um assunto que ainda apresenta muitas questões em aberto, além de uma grande variedade de produtos e serviços oferecidos com essa finalidade. Esta seção abrange um pouco das principais tecnologias e modelos ligados à Computação em Nuvem.

2.2.1 Modelos Relacionados

2.2.1.1 Sistemas Distribuídos

Como mencionado anteriormente, o termo nuvem remete a uma abstração de toda complexidade computacional subjacente capaz de oferecer recursos virtualmente ilimitados e de forma "elástica". Entretanto, para o usuário final alocar e desalocar recursos infinitamente, os provedores de nuvem contam com *data centers* que, unidos, tornam possível a elasticidade e a ideia de recurso infindo, constituindo as principais características da Computação em Nuvem. A forma com a qual esses *data centers* se comunicam trocando mensagens entre si através de uma rede de computadores, é um exemplo de um sistema computacional distribuído.

Um sistema distribuído pode ser definido como um conjunto de componentes, localizados em computadores, que se comunicam trocando mensagens entre si com o intuito de coordenar suas ações. Os computadores interligados por meio de uma rede podem estar separados por qualquer distância. Eles podem estar em continentes separados, no mesmo prédio ou na mesma sala (COULOURIS et al., 2005).

A Figura 2 ilustra um sistema distribuído típico, no qual a camada de middleware é composta pelo *software* responsável pelo provimento de uma interface de comunicação única e pela coordenação de usuários e aplicações para a utilização de máquinas distintas, cada uma com seu próprio sistema operacional.

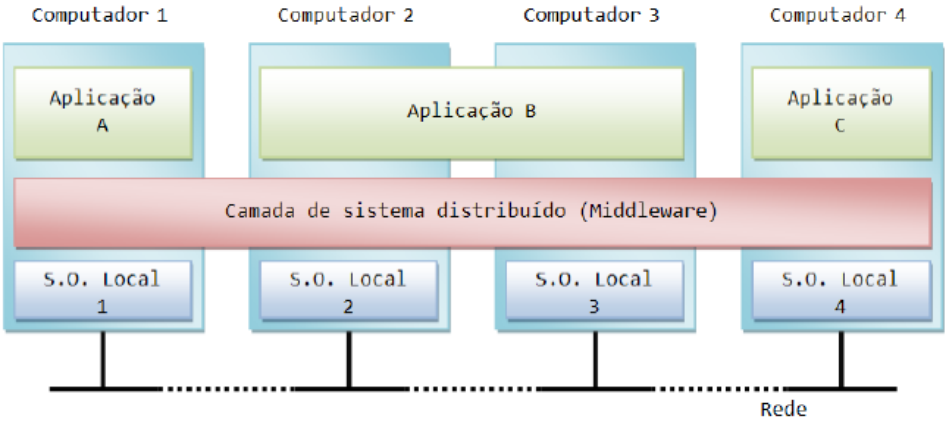

Figura 2: Sistema distribuído (TANENBAUM, 2006)

Assim, é lógico afirmar que o significado de nuvem faz referência a um grande sistema distribuído responsável por todo o compartilhamento de recursos disponíveis.

2.2.1.2 Grades Computacionais

Muitas das características encontradas em grades computacionais também são encontradas na Computação em Nuvem. Isso é devido aos objetivos em comum em ambos os modelos, tais como: redução dos custos computacionais, compartilhamento de recursos e aumento de flexibilidade e confiabilidade (VERDI et al., 2010). No entanto, existem algumas diferenças que precisam ser enfatizadas. O trabalho Vaquero et al. (2009) realiza um estudo das diferentes características associadas com Computação em Nuvem e as compara com as características associadas com grades computacionais. Abaixo serão elencados alguns itens comparativos.

Modelo de pagamento e origens: as grades computacionais surgiram através de financiamento público, na maioria das vezes patrocinadas por projetos dentro de universidades. O modelo *Cloud Computing* é motivado por aspectos comerciais onde grandes empresas criam estratégias de mercado com interesses nos lucros. Tipicamente, os serviços em grade são

cobrados usando uma taxa fixa por serviço, enquanto que os usuários dos serviços oferecidos nas *clouds* são cobrados pelo modelo *pay-per-use*. Muitas aplicações não usam a mesma capacidade computacional de armazenamento e recursos de rede. Sendo assim, o modelo de cobrança deve considerar o pagamento separado para cada tipo de recurso utilizado;

Compartilhamento de recursos: as grades computacionais compartilham os recursos entre as organizações usuárias através do modelo "mais justo possível". A Computação em Nuvem fornece a quantidade de recursos desejados para cada usuário dando a impressão de recurso dedicado. Esta noção de recurso dedicado é possível através do uso de virtualização, aspecto ainda pouco explorado pelas grades;

Virtualização: as grades computacionais usam interfaces para esconder a heterogeneidade dos recursos computacionais. A virtualização utilizada em grades computacionais é ainda muito simplista. A virtualização utilizada em *Cloud Computing* ocorre de forma plena, possibilitando que usuários instalem máquinas virtuais e sistemas operacionais específicos nestas máquinas virtuais. A migração/mobilidade de máquinas virtuais também é um aspecto comum dentro da nuvem e permite a otimização do uso de recursos de energia e resfriamento;

Escalabilidade e gerenciamento: a escalabilidade em grades ocorre através do aumento no número de nós de processamento. A escalabilidade em *Cloud Computing* ocorre através de um redimensionamento do hardware virtualizado. O gerenciamento das grades computacionais é dificultado pois não há tipicamente uma única entidade proprietária de todo o sistema. Por outro lado, as *clouds* encontradas atualmente são controladas por uma única entidade administrativa, muito embora exista uma tendência em se criar federações de nuvens;

Padronização: a maturidade das grades computacionais fez com que vários fóruns fossem criados para a definição de padronização. Neste sentido, esforços para padronização de interfaces para os usuários assim como padronização de interfaces internas alavancaram a interoperabilidade de grades computacionais. Em *Cloud Computing*, as interfaces de acesso aos serviços são muito parecidas com as interfaces das grades, entretanto, as interfaces internas são proprietárias e dificultam a criação de federação de nuvens. Atualmente há várias iniciativas para definição de padrões para Computação em Nuvem. Um dos desafios principais é a padronização do formato das imagens virtuais e APIs de migração.

De um modo geral, grade computacional se refere ao processamento distribuído e paralelo, ou seja, as tarefas são quebradas em várias partes e distribuídas nos nós da grade para o processamento, e então as partes são unidas para se obter o resultado. Enquanto que na Com-

putação em Nuvem, os recursos necessários são adquiridos para a realização de uma tarefa computacional em um determinado tempo (VERDI et al., 2010).

Conforme mencionado anteriormente, a Computação em Nuvem e as grades computacionais possuem algumas características em comum. Desta forma, pode-se dizer que a *cloud computing* segue os princípios do modelo de Computação em Grade.

2.2.1.3 Computação utilitária

Na *Utility Computing*, os usuários têm os recursos envolvidos monitorados, de forma clara, para as partes envolvidas, provedor e usuário, e só serão tarifados pela quantidade e qualidade daquilo que utilizarem. Usuários não precisam se preocupar com escalabilidade, pois a capacidade de armazenamento é infinita; terão disponibilidade total, isto é, podem ler e gravar a qualquer tempo, sem nunca serem bloqueados; e nem se preocuparem com *backups*, pois se os componentes falharem, o provedor é responsável por substituí-los (BRANTNER et al., 2008).

O objetivo da *Utility Computing* é fornecer componentes básicos como armazenamento, processamento e largura de banda de uma rede como uma "mercadoria" através de provedores especializados com um baixo custo por unidade utilizada (SOUSA et al., 2009).

A Computação em Nuvem é uma evolução dos serviços e produtos de tecnologia da informação sob demanda, chamada de *Utility Computing* (BRANTNER et al., 2008).

2.2.2 Principais tecnologias

2.2.2.1 Virtualização

Um maior investimento em ambientes de Computação em Nuvem pode consistir em um aumento do número de usuários do provedor. Porém, nada garante que os servidores não se tornem subutilizados. Mesmo que todos servidores estejam ligados, executando tarefas dos usuários, é possível que o compartilhamento de recursos esteja bem abaixo do esperado. A utilização mais eficiente desses servidores, busca um maior retorno no investimento feito no hardware. Maior eficiência significa mais trabalho obtido pelo mesmo hardware. Isso é obtido através da distribuição de seus recursos (espaço em memória principal, processador, espaço em disco, etc) entre diferentes programas (CARISSIMI, 2008).

A virtualização é uma tecnologia-chave na Computação em Nuvem, a qual permite que vários sistemas operacionais executem no mesmo *hardware* físico, compartilhando memória, armazenamento e processamento, de tal forma que os serviços em execução nas MVs sejam executados separadamente de todos os outros processos nas MFs (CARISSIMI, 2009).

A utilização de técnicas de virtualização pode trazer outros benefícios, como (VMWARE,

2013):

- Redução dos custos operacionais;
- Redução do tempo gasto em tarefas administrativas de TI;
- Facilidade para fazer backup e proteção de dados;
- Consolidação de servidores;
- Aumento da disponibilidade de aplicativos;
- Facilidade para recuperação de falhas.

A implementação de máquinas virtuais ou VMM pode ser obtida através de duas técnicas (CARISSIMI, 2008):

Virtualização total Consiste em prover uma réplica (virtual) do hardware subjacente de tal forma que o sistema operacional e as aplicações podem executar como se tivessem executando diretamente sobre o hardware original. A grande vantagem dessa abordagem é que o sistema operacional hóspede[2] não precisa ser modificado para executar sobre a VMM. Porém, todas as instruções executadas pelo sistema hóspede devem ser testadas na VMM, o que representa um custo de processamento.

Para-virtualização Nessa abordagem, a VMM permite que o sistema hóspede acesse alguns recursos do *hardware* diretamente, sem a intermediação do mesmo. Em outras palavras, o acesso ao hardware é apenas monitorado pela VMM, que informa ao sistema hóspede seus limites, como as áreas de memória e de disco disponíveis

A virtualização, como peça chave na construção de um ambiente computacional em nuvem, constitui uma das já mencionadas características da *Cloud Computing*, chamadada "elasticidade", que é obtida através da criação de novas instâncias de máquinas virtuais quando se faz necessário, oferecendo ao cliente a capacidade de alocar e desalocar recursos deliberadamente.

2.2.2.2 Web Services

Para o fornecimento de serviços de uma maneira mais rica e mais estruturada de interoperabilidade entre servidores é necessário o uso de *web services*. Os *web services* disponibilizam uma interface de serviço que permite aos servidores interagirem uns com os outros.

[2]Processo ou sistema que executa sobre uma máquina virtual.

Eles fornecem uma base por meio da qual um servidor em uma organização pode interagir com outro servidor em outra organização, sem supervisão humana. Em particular, os serviços *web* fornecem serviços que integram vários outros serviços.(COULOURIS et al., 2005).

Para a representação externa dos dados, na comunicação entre clientes e serviços *web*, é usada como padrão a linguagem de representação textual XML (Linguagem de Marcação Extensível) . O responsável pela especificação de regras de uso do XML para empacotar mensagens, por exemplo, para suportar um protocolo de requisição e resposta, é o SOAP.

A Computação em Nuvem faz com que seja possível a utilização de infraestrutura de *hardware* e *software* remotamente, oferecendo-os como serviços aos usuários finais. Isto é possível devido à utilização de interfaces de *web services*, através das quais requisições são recebidas pelos servidores de gerenciamento dos provedores, que administram o provimento de recursos de acordo com suas políticas internas de segurança e SLA estabelecidos com seus clientes.

2.2.2.3 Middleware

O *middleware* é uma parte essencial para a Computação em Nuvem, pois sem ele a comunicação entre máquinas distintas, utilizando diferentes representações de dados para a comunicação, não seria possível. Ele pode ser definido como uma camada de *software* cujo objetivo é mascarar a heterogeneidade e fornecer um modelo de programação conveniente para os programadores de aplicativos (COULOURIS et al., 2005).

Um *middleware* é composto por um conjunto de processos ou objetos, em um grupo de computadores, que interagem entre si de forma a implementar comunicação e oferecer suporte para compartilhamento de recursos e aplicativos distribuídos. Em particular, ele simplifica as atividades de comunicação de programas aplicativos por meio do suporte de abstrações como a invocação a métodos remotos, a comunicação entre um grupo de processos, a notificação de eventos, o particionamento, posicionamento e recuperação de dados compartilhados entre computadores, a replicação de objetos de dados compartilhados e a transmissão de dados multimídia em tempo real (COULOURIS et al., 2005).

Visto as definições das seções anteriores, pode-se afirmar que a Computação em Nuvem, os sistemas distribuídos e os *middlewares* estão intrinsecamente ligados. Porém, vale ressaltar que existem incentivos, tanto por parte da academia como por parte do mercado, para o aprimoramento de *middlewares* e sistemas de gerenciamento como um todo. Mais uma vez, há diferentes tecnologias sendo usadas em diferentes provedores de nuvem, portanto, não existe um padrão.

2.3 Modelos de Serviços

O novo paradigma proposto pela *Cloud Computing* baseia-se na entrega de uma vasta variedade de recursos através de serviços disponíveis em escala mundial. Isto implica em um impacto profundo na forma como empresas e pessoas utilizam o *hardware* e o *software* de que necessitam, além de elevar ainda mais a importância das redes de comunicação através das quais esses recursos se tornam acessíveis. A variedade de serviços disponíveis em uma nuvem faz com que sua classificação divirja sobremaneira. Nomenclaturas como CaaS (Comunicação como Serviço), MaaS (Gerenciamento como Serviço), DaaS (Dados como Serviço), ITaaS (TI como Serviço) e até mesmo XaaS (Tudo como Serviço) são encontradas com relativa facilidade na literatura (PALLIS, 2010). Todavia, a estrutura da Computação em Nuvem é visivelmente formada por três camadas fundamentais (STANOEVSKA-SLABEVA, 2010). Logicamente interconectados (Figura 3), os elementos que compõem esta visão triádica da nuvem são a IaaS (Infraestrutura como Serviço), a PaaS (Plataforma como Serviço) e o SaaS (Software como Serviço).

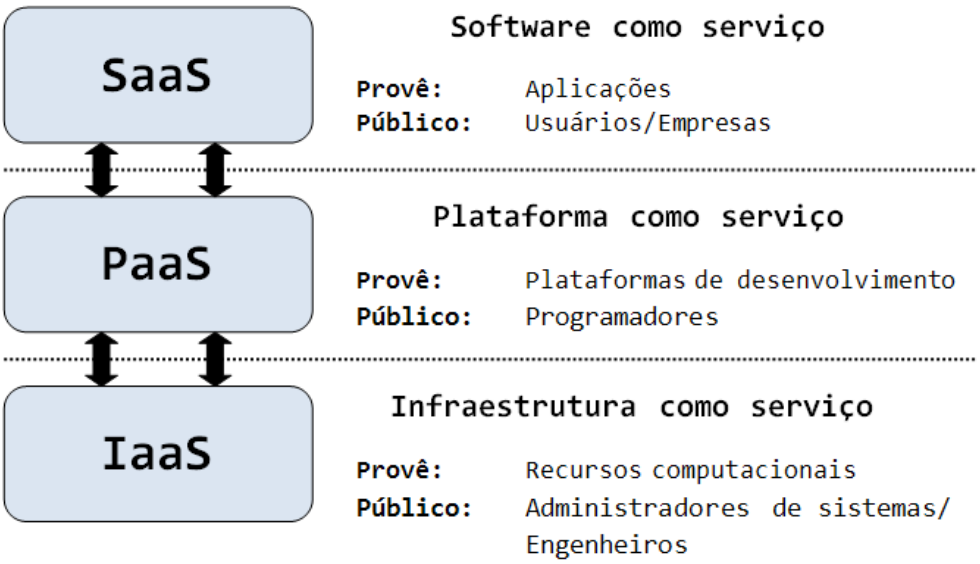

Figura 3: Visão em camadas dos modelos de serviço da Computação em Nuvem. Adaptada de STANOEVSKA-SLABEVA (2010)

Atualmente, este tipo de classificação é amplamente aceito no mercado e na academia, pois engloba grande parte dos subtipos encontrados e define suas funcionalidades de forma coerente e concisa.

2.3.1 Software como Serviço

A forma tradicional de utilização de um programa se faz através da sua instalação e execução diretamente no computador do usuário. Já o modelo de SaaS consiste no provimento

via rede de programas executados em sistemas e infraestruturas de terceiros, através de uma técnica de cobrança baseada na taxa de utilização do usuário.

Comumente chamada de *pay-per-use*, essa técnica permite que clientes eliminem grande parte dos custos atrelados ao uso de um determinado *software*, excetuando-se os relativos à utilização das redes de comunicação. Não obstante, o emprego de SaaS gera uma grande dependência do usuário em relação a recursos que se encontram além de seu domínio, o que levanta muitos questionamentos a respeito da adesão ao modelo.

2.3.2 Plataforma como Serviço

A camada de Plataforma como Serviço (PaaS) é direcionada aos programadores, oferecendo-os um ambiente autoconfigurável através do qual é possível o desenvolvimento e o teste de aplicações sem a preocupação com a infraestrutura necessária para tal. Por meio do modelo de PaaS, o programador pode criar e executar uma aplicação inteiramente na nuvem. As variações no uso da aplicação são então gerenciadas de forma integral pela plataforma. Caso seu nível de utilização cresça, um acréscimo na quantidade de recursos alocados é efetuado (RIGHTSCALE, 2010). Uma ação análoga é realizada nos casos de decréscimo do nível de utilização.

2.3.3 Infraestrutura como Serviço

Em contrapartida à venda direta de componentes de hardware e servidores completos, o modelo de Infraestrutura como Serviço (IaaS) oferece ao consumidor poder de processamento, armazenamento de dados, estrutura de rede e uma série de outros recursos básicos de forma virtualizada e seguindo os preceitos do self-service sob demanda, elasticidade e multi-arrendamento. Instituições que aderem a este tipo de serviço não têm controle direto sobre a infraestrutura física, mas podem utilizar os recursos virtualizados que lhe são oferecidos da forma que lhes convier, salvo restrições estabelecidas pelo provedor.

Ao eliminar a necessidade de posse de uma infraestrutura particular, o emprego de IaaS têm impacto direto em custos relacionados ao consumo de energia, espaço físico e manutenção de hardware.

2.4 Modelos de Implantação

Na seção 2.3, foi feita uma explanação no que tange os serviços oferecidos pela Computação em Nuvem. Além disso, o NIST classifica como os ambientes computacionais em nuvem são implantados, de uma maneira a discernir o acesso aos usuários em geral. Os principais tipos de nuvem podem ser classificados como: privada, pública, híbrida e comunitária (BADGER et al., 2011).

2.4.1 Nuvem Pública

No modelo de implantação de nuvem pública, a infraestrutura de nuvem é disponibilizada para o público em geral no modo "pay-per-use", sendo acessada por qualquer usuário que conheça a localização do serviço (VERDI et al., 2010).

2.4.2 Nuvem Privada

Compreende uma infraestrutura de nuvem operada unicamente por uma organização. Os serviços são oferecidos para serem utilizados internamente pela própria organização, não estando disponíveis publicamente para uso geral (VERDI et al., 2010).

2.4.3 Nuvem Comunitária

Neste modelo, ocorre o compartilhamento de uma nuvem entre várias organizações com interesses em comum, tais como segurança e política. Este tipo de modelo de implantação pode existir localmente ou remotamente e geralmente é administrado por alguma organização da comunidade ou por terceiros (SOUSA et al., 2009).

2.4.4 Nuvem Híbrida

As nuvens híbridas são uma combinação de duas ou mais nuvens distintas. Apesar de manterem-se como entidades únicas, as nuvens se comunicam através de interfaces padronizadas, criando-se a possibilidade de compartilhamento de recursos e aplicações. Muitas organizações empregam este modo de implantação para tirar proveito dos benefícios trazidos pelas nuvens públicas, enquanto mantêm algumas aplicações e dados em seu domínio administrativo através de nuvens privadas, evitando, desta forma, problemas relacionados à segurança de suas informações (SUN, 2010).

2.5 Infraestrutura

A *Cloud Computing* vem se tornando uma grande aliada de empresas que oferecem serviços por meio da TI. Porém, ela impõe desafios técnicos para suportar uma crescente demanda por serviços de *cloud*. Alguns dos principais desafios ligados aos *data centers* tradicionais podem ser: o alto consumo de energia elétrica e um maior retorno no investimento feito no *hardware*.

Esta seção discute, principalmente, os desafios e as limitações da infraestrutura tradicional, tais como os atuais sistemas de gerenciamento de recursos.

2.5.1 Classificação

A infraestrutura da nuvem é formada pelos *data centers* que abrigam os servidores que, mediante diferentes níveis de organização e técnicas de virtualização, oferecem os serviços em nuvem. Portanto, os *data center* são a manifestação física da Computação em Nuvem, sendo a infraestrutura de rede a base de comunicações que habilita o paradigma de *Cloud Computing*, interligando servidores físicos em grande escala. Dependendo do tamanho da própria infraestrutura física e sua localização, *os data centers* podem ser classificados como mega, micro, nano ou baseados em contêineres (VERDI et al., 2010):

Mega data centers: nos mega *data centers*, encontram-se milhares de servidores interconectados consumindo um alto índice de energia. Há vários fatores a serem observados na construção desses enormes "armazéns" de servidores. Como (1) energia, deve-se preferir a escolha de locais com energia barata e abundante e (2) uma boa conectividade à internet. Os mega *data centers* atendem principalmente àquelas aplicações que requerem grandes quantidades de CPU , memória RAM e armazenamento.

Micro data centers: os micro *data centers* suportam principalmente aplicações altamente interativas, como por exemplo, aplicações de escritório e *e-mail*. As principais aplicações que são oferecidas por esses data centers são aquelas que não necessitam de troca de grandes volumes de dados. Geralmente são usados como servidores front-ends com rápidas respostas ao usuários;

Nano data centers: o conceito principal deste tipo de *data center* surge do paradigma P2P. Nesse cenário, os equipamentos dos usuários final são considerados uma extensão natural do *data center*, que serviços e armazenamento de dados podem ser providos. O lado positivo desta adoção de data center, é que os custos são reduzidos para o provedor da infraestrutura de nuvem. Po outro lado, surgem inúmeros desafios, como garantias de desempenho, confiabilidade e segurança.

Data centers baseados em contêineres: é visto como um modelo interessante para a adoção de *data centers*. Os contêineres podem caber até 2000 servidores, formando um bloco computacional modularizado (ver figura 4). É uma excelente opção quando se quer implantar *data centers* temporários, como em eventos esportivos ou culturais, por exemplo. Outros pontos positivos acerca desse modelo, é que os contêineres permitem ser instalados sob-demanda e podem economizar cerca de 40% a 50% das despesas operacionais de *data center* tradicionais.

Figura 4: Um modelo de *data center* baseado em contêineres. (VERDI et al., 2010)

2.5.2 Arquitetura tradicional

Independentemente de como o *data center* se classifica, a infraestrutura é formada por milhares de servidores, cada qual com suas redes de comunicação, armazenamento, distribuição de energia e sistemas de refrigeração. Existe também uma preocupação com a eficiência dos custos, o fator número um nas decisões de projeto, implementação e com a procura de novas abordagens na construção e operação desses sistemas (VERDI et al., 2010).

A Figura 5 serve como referência de projetos para *data centers*. Ela emprega a divisão da infraestrutura utilizando a abordagem de topologia em camadas para prover escalabilidade, alto desempenho, flexibilidade, resiliência e facilidade de manutenção (VERDI et al., 2010).

Sucintamente, a ideia da Figura pode ser explicada da seguinte forma: o *Core* é o terceiro nível de *switches IP/Ethernet* e é categorizado como *switches*/roteadores de *Core*; *Aggregation* é onde ficam os *switches IP/Ethernet* de agregação, também conhecidos como *switches End-of-Row*, que agregam os *clusters* de servidores. Este segundo nível de domínio de comutação pode, potencialmente, abranger mais de dez mil servidores individuais; *Services* são módulos integrados nos *switches* de agregação ou disponibilizados em equipamentos dedicados, que proveem serviços tais como balanceadores de carga, roteadores de conteúdo, segurança (*firewall*, detecção de intrusão, proteção contra ataques de DDoS), análise e monitoramento da rede, etc; A camada *Access* é constituida por servidores físicos que são tipicamente montados em conjunto dentro de um rack e interligados através de um *switch Ethernet* de acesso. Este *switch* é denominado *Top-of-Rack* e usa interfaces de *uplink* de 1 ou 10 Gbps para se interligar com um ou mais switches *IP/Ethernet* de agregação.

Segundo Verdi et al. (2010), um *data center* suporta dois tipos de tráfego: o tráfego que entra e sai do *data center* e o tráfego que é gerado e flui apenas internamente ao *data center*. Por exemplo, em aplicações de busca (*searching*) o tráfego interno domina devido à necessidade de realizar indexações e sincronizações. Por outro lado, em aplicações de vídeo sob-demanda, o tráfego externo prevalece.

O tráfego vindo de fora, ou seja, vindo da internet, é distribuído para um conjunto de servidores responsáveis pelo processamento. Normalmente, as comunicações com os usuários externos são atendidas por servidores denominados de *Front-End*. Para completar o serviço, es-

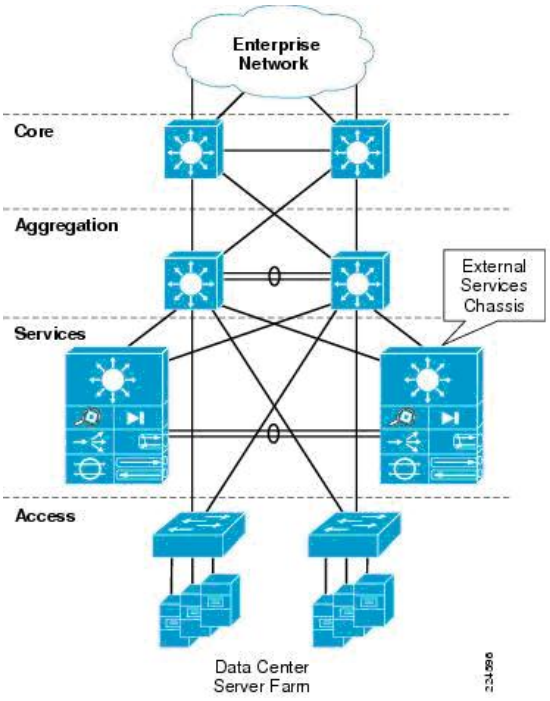

Figura 5: Topologia em camadas de um *data center*. (VERDI et al., 2010)

tes servidores *Web Front-Ends* tipicamente se comunicam com outros servidores denominados *Back-Ends* e esses, posteriormente, acessam os meios de armazenamento distribuídos. A Figura 6 ilustra perfeitamente.

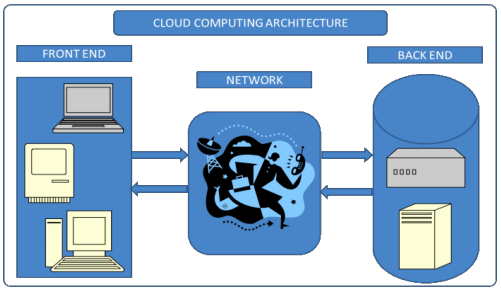

Figura 6: Servidores Back-end e Front-end

Segundo os estudos realizados em Verdi et al. (2010), as arquiteturas tradicionais de *data centers* apresentam limitações consideráveis. Duas delas podem ser resumidas da seguinte forma:

Escalabilidade os protocolos atuais de roteamento, encaminhamento e gerenciamento não oferecem a escalabilidade necessária para atingir a ordem de grandeza necessária nos *data centers*. As soluções atuais de camada 2 utilizam broadcast, criando um cenário não escalável devido ao elevado nível de sinalização. Por outro lado, as soluções de camada 3 exigem a configuração dos equipamentos, seja na definição de sub-redes nas quais os

equipamentos estão incluídos ou mesmo na sincronização de servidores DHCP para efetuar a correta atribuição de endereços.

Eficiência Energética os *data centers* tradicionais usam mecanismos básicos de refrigeração seguindo a premissa de que se o *data center* cresce, instala-se mais equipamentos de refrigeração, causando um impacto significativo no consumo de energia mensal. Além dos aspectos de refrigeração e sistemas de distribuição da energia na infraestrutura, ainda há uma margem alta de eficiência energética que não é atingida pelos servidores e os equipamentos de rede atuais. Especialmente em relação à proporcionalidade do consumo de energia em função da carga de trabalho pois, nos projetos atuais, o consumo dos equipamentos (e da rede como um todo) trabalhando com uma carga mínima, não consomem proporcionalmente menos energia do que trabalhando ao máximo da sua capacidade.

Para ter-se uma ideia da distribuição dos custos com a infraestrutura de TI tradicional, levando em conta o baixo índice de eficiência energética, os principais custos de um *data center* da ordem de 50.000 servidores construído seguindo as melhores práticas da indústria (qualidade, alta disponibilidade, etc.), podem ser estruturados conforme a figura .

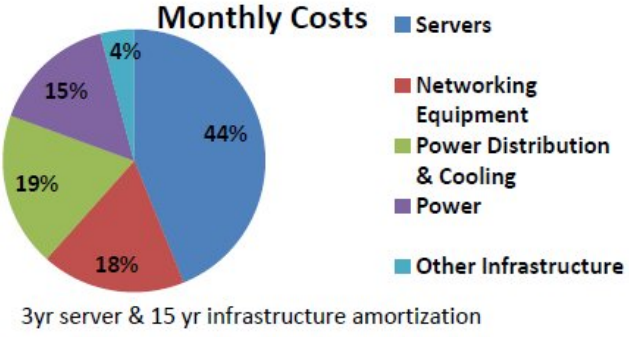

Figura 7: Distribuição dos custos mensais em um *data center* com 50 mil servidores. Extraída de Hamilton (2008)

Ao observar-se o gráfico acima, constata-se que 62% do custo pertence à infraestrutura de TI (44% em servidores e 18% em equipamentos de rede). Os números específicos podem ser discutidos e variar de um caso para outro. Porém, levando em conta a tendência atual relativa ao aumento dos custos com energia e infraestrutura, a conclusão é que os custos totais associados à energia (soma dos custos com a energia consumida, para efetuar a refrigeração e a infraestrutura necessária para distribuição de energia) são hoje comparáveis aos custos dos equipamentos de TI e poderão dominar os custos de um *data center* (VERDI et al., 2010).

Conclui-se então, que as técnicas convencionais não oferecem o desempenho nem a agilidade necessária. Entende-se por agilidade a capacidade de realocar dinamicamente servidores entre os serviços oferecidos pela nuvem com o propósito de otimizar a alocação da carga de trabalho (GREENBERG et al., 2009).

Nesse sentido, o uso das práticas de Computação Verde é visto como uma excelente opção a se aplicar nos projetos de *data centers*. Tendo em mente os dois desafios supracitados, a Computação Verde pode contribuir fortemente para amenizar os gastos com energia elétrica e favorecer ao desenvolvimento sustentável do meio-ambiente.

2.6 Computação Verde

O uso intensivo de recursos, como aplicações científicas, cria uma demanda crescente por infraestrutura computacional de alto desempenho. Isto levou a construção de grandes *data centers* que consomem uma enorme quantidade de energia elétrica. Apesar das melhorias em eficiência energética dos *hardwares* atuais, o consumo energético continua a crescer devido ao aumento das necessidades de recursos computacionais. Por exemplo, em 2006 o custo do consumo energético gerado pelas infraestruturas de TI nos EUA foi de aproximadamente 4,5 bilhões de dólares e foi provavelmente o dobro em 2011 (BELOGLAZOV; BUYYA, 2010b).

Os *data centers* não são apenas caros. Eles agridem o meio-ambiente emitindo uma grande quantidade de gás carbônico, em uma taxa igualmente perniciosa de ambos os países, Argentina e Holanda (BELOGLAZOV; BUYYA, 2010b). Devido aos problemas, foi necessário repensar a abordagem utilizada nos projetos convencionais de *data centers*, resultando no termo que hoje é chamado de Computação Verde.

Computação Verde, também conhecida como *Green Computing*, ou ainda TI verde, se refere ao uso eficiente de recursos computacionais, minimizando o impacto ambiental, maximizando sua viabilidade econômica e assegurando os deveres sociais. Através dela pode-se dizer que no futuro as ações tecnológicas irão prejudicar o mínimo possível o meio ambiente e serão mais sustentáveis. Quando discutido sobre TI verde, o consumo de energia sempre vem em primeiro lugar, pois é o fator que mais gera despesas dentre os tópicos pesquisados nesta área (SHULZ, 2004).

2.6.1 Funcionamento

Abaixo são mostradas as características da Computação em Nuvem Verde (WERNER, 2011):

Flexibilidade: Oferece as mesmas características de reconfiguração da Computação em Nuvem, além de poder gerenciar o status das máquinas físicas (ligando/desligando) quando necessário, assim como possibilita o agrupamento dos recursos, e possibilita a mobilidade da estrutura;

Disponibilidade: Oferece as mesmas características da Computação em Nuvem, e poderia ge-

renciar as máquinas físicas (hibernando) e remover máquinas virtuais ociosas;

Custo: Com a funcionalidade de movimentação de máquinas virtuais extra nuvem, *data centers* adotam uma configuração minimalista, impactando também no aumento da vida útil dos equipamentos. Já a estrutura diminui os seus custos mensais, reduzindo o consumo de energia derivado das políticas de "desligamento e hibernação";

Sustentabilidade: Com a funcionalidade de movimentação de máquinas virtuais, tem-se a possibilidade de, em períodos de baixa demanda, concentrar o processamento das máquinas virtuais em poucos servidores físicos permitindo o desligamento dos equipamentos ociosos. O sistema trabalha em conjunto ao ambiente, inferindo a estratégia de trabalho sob demanda também aos equipamentos externos (por exemplo, refrigeração e rede), assim como leva em consideração os fatores externos, como agir proativamente em caso de desastre (por exemplo, incêndio).

2.6.2 Trabalhos Relacionados

Contextualizando o atual cenário da Computação em Nuvem Verde, pode-se notar que um dos principais desafios ligados ao gerenciamento dos *data centers* é a amortização dos custos energéticos. O presente trabalho e todos os outros citados nesta seção, possuem em comum o objetivo de contribuir com a Computação Verde.

É notável que a busca por soluções que atenuem os custos é algo pertinente nos projetos de infraestrutura e em trabalhados relacionados ao assunto. Barroso e Hölzle (2007) sugerem componentes verdes que atenuem o consumo de energia em função da carga e Qureshi et al. (2009) propõem o encaminhamento de pacotes para *data centers* onde a energia é mais barata.

Em Lago et al. (2011), foi projetado e avaliado um algoritmo para escalonamento de tarefas em nuvens, usando o conceito de *Computação Verde*. O algoritmo focou na minimização do consumo de energia em *data centers* e, consequentemente, na redução da poluição potencial gerada. Para atingir esse objetivo foram usados princípios de gerência distribuída de energia, ajuste dinâmico de tensão e frequência, migração de máquinas virtuais e desligamento de servidores subutilizados.

No artigo de Werner et al. (2012), é proposto um sistema de gerenciamento de recursos utilizando os princípios da Teoria das Organizações no intuito de classificar os serviços de uma nuvem. A gestão da Nuvem por meio dos princípios da Teoria da Organização fornece a possibilidade de configuração automática do Sistema de Gerenciamento, uma vez que a adição de um novo elemento (por exemplo, Máquina virtual, Máquina física) é apenas uma questão de incluir um novo serviço no Grupo de Gestão.

Já em Lago et al. (2012) é proposto um algoritmo com qualidade de serviço baseado

em prioridades de cargas de trabalho. O algoritmo faz com que as cargas de trabalho com maior prioridade sejam executadas primeiro que as demais, mantendo o mesmo consumo de energia e, consequentemente, aumentando a qualidade de serviço. Os trabalhos Beloglazov e Buyya (2010b), Binder e Suri (2009) realizam estudos na tentativa de colaborar com maneiras eficientes de diminuir o custo operacional e aumentar a qualidade de serviço em *data centers*.

2.7 Gerenciamento de Recursos

Os sistemas de gerenciamento são fundamentais para o funcionamento da Computação em Nuvem. Eles são responsáveis por controlar e monitorar os recursos para cada tipo de serviço oferecido ao usuário final (armazenamento, processamento e largura de banda) (VERDI et al., 2010). Em outras palavras, a consolidação dinâmica de servidores, o desligamento de máquinas virtuais subutilizadas, o cumprimento dos SLAs, seriam uma tarefa dificilmente empregada com a ausência de um sistema de gerência robusto. Nesse sentido, o que inquieta o mercado e a academia, e que também tem sido foco de pesquisas por ambas as partes, é o desenvolvimento de sistemas consistentes capazes de gerenciar recursos de uma maneira a qual suportem o consumo eficiente de energia.

A consolidação de hardware e a redução de redundância podem alcançar a eficiência energética. Máquinas virtuais podem ser movidas, copiadas, criadas e deletadas dependendo das decisões de gerenciamento. Assim sendo, os servidores subutilizados podem ser desligados ou "hibernados" para economizar energia (ANDREAS et al., 2009).

Resumindo, o papel dos sistemas de gerência é intermediar as cargas de trabalho provenientes dos usuários, para só depois, encaminhá-las aos servidores no *data center*. A Figura 8 ilustra uma arquitetura que suporta a alocação de serviços com eficiência energética, e pode facilitar o entendimento de como tais sistemas atuam.

De acordo com Beloglazov e Buyya (2010a), nessa arquitetura existem basicamente quatro entidades envolvidas . São elas:

Consumidores/Brokers: Consumidores da nuvem (ou *brokers*) enviam requisições de serviços de qualquer lugar do mundo. É importante ressaltar que pode haver uma diferença entre consumidores e usuários de nuvem. Por exemplo, um consumidor pode ser uma empresa, que envia cargas de trabalho de acordo com o seu número de usuários que estão acessando à nuvem.

Gerenciador Verde de Recursos: Age como a interface entre a infraestrutura da nuvem e os consumidores e requer a interação dos seguintes componentes para suportar o gerenciamento de recurso com o consumo de energia eficiente.

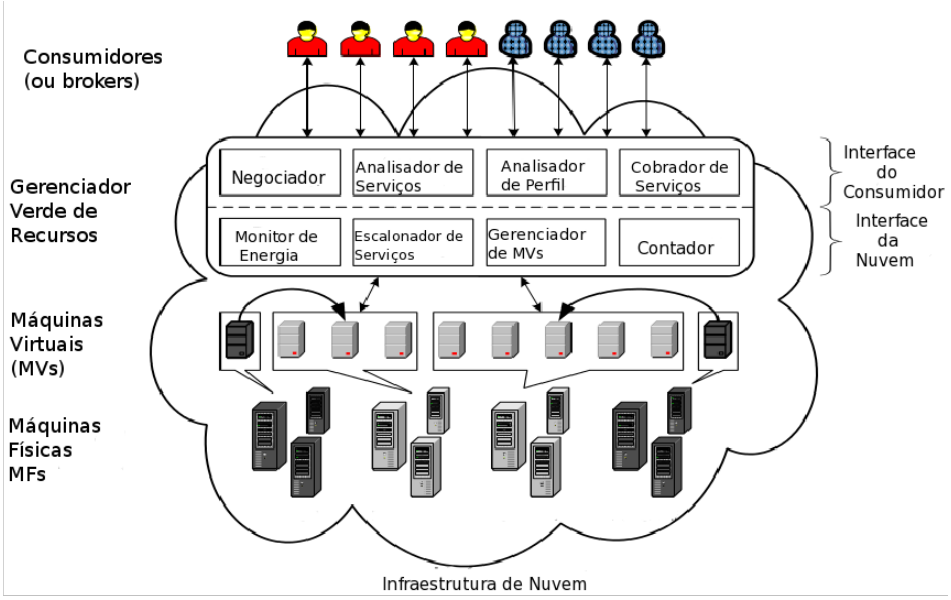

Figura 8: Arquitetura de alto nível de um sistema de gerência. Adaptada de Beloglazov e Buyya (2010a)

- Negociador: Negocia o SLA com os consumidores a respeito de preços e penalidades entre o provedor e os mesmos, dependendo das condições de QoS. Exemplificando, aplicações *Web* podem ter uma métrica de QoS onde 95% das requisições são servidas em menos de três segundos.
- Analisador de Serviços: Analisa os serviços desejados de uma requisição antes de decidir se vai rejeitá-la ou aceitá-la. Além disso, esse componente detém algumas informações, como o consumo energético e a sobrecarga das máquinas virtuais e físicas, cedidas pelo Monitor de Energia e pelo Gerenciador de MVs , respectivamente.
- Analisador de Perfil: Reuni informações específicas de cada consumidor, e determina a prioridade de acordo com o perfil de cada um.
- Cobrador de Serviços: Gerencia o fornecimento e a demanda por serviços computacionais (e como são cobrados) e também facilita a prioridade de alocação de serviços de maneira efetiva.
- Monitor de Energia: Observa e determina qual máquina física deve ser ligada ou desligada.
- Escalonador de Serviços: Determina os direitos de acesso das máquinas virtuais aos recursos físicos. Por exemplo, qual MV fica com a maior parte de armazenamento, processamento, etc. Além disso, decide quando as MVs deverão ser adicionadas ou removidas para atender a demanda.
- Gerenciador de MVs: Mantém o controle da disponibilidade de máquinas virtuais

e seus privilégios de recursos. Também é responsável pela migração de máquinas virtuais entre máquinas físicas.

- Contador: Mantém o uso atual dos recursos e computa os custos. As informações sobre o histórico de uso também podem ser usadas para melhorar as decisões de alocação de serviços.

MVs: As MVs podem ser dinamicamente iniciadas e paradas em uma única máquina física de acordo com a demanda de requisições, provendo a máxima flexibilidade ao configurar várias partições de recursos computacionais no mesmo nó físico. Desta forma, atendendo à demanda por recursos específicos em um mesmo servidor físico. MVs podem executar aplicações baseadas em diferentes sistemas operacionais mesmo estando em uma mesma máquina física. Adicionalmente, migrando MVs entre máquinas físicas, cargas de trabalho podem ser consolidadas e recursos subutilizados podem ser "hibernados", desligados ou configurados para operar em um nível de baixo desempenho (por exemplo, usando DVFS), desse modo economizando energia.

Máquinas Físicas: Os servidores computacionais subjacentes proveem a infraestrutura de *hardware* necessária para criar recursos virtualizados para atender à demanda de serviços.

O que se pode notar na arquitetura ilustrada pela Figura 8 é que a parte essencial para um gerenciamento eficiente de energia, é ter um conjunto de componentes que constituam o Gerenciador Verde de Recursos. Resumidamente, esses componentes têm a tarefa de tratar as MVs como unidades de escalonamento para serem alocadas em recursos físicos heterogêneos, privilegiando principalmente o baixo consumo de energia. Como foi discutido nas seções anteriores, o consumo energético apresenta um alto custo e é tido como um dos principais desafios da Computação em Nuvem. Desta forma, é reforçada mais ainda, a ideia de que é preciso melhorar o funcionamento dos sistemas de gerenciamento, dando-os maior inteligencia na tomada de decisões. Mas para isso, os projetos de infraestrutura devem adotar, cada vez mais, novos mecanismos que colaborem para a eficiência dos sistemas de gerenciamento.

2.8 Resumo

Este capítulo apresentou os principais aspectos relacionados ao paradigma da Computação em Nuvem, possibilitando o entendimento dos conceitos relacionados a estes trabalho. Inicialmente, foi feita uma breve definição do termo. Em seguida, os modelos computacionais relacionados e as principais tecnologias envolvidas foram discutidos. Posteriormente, foi apresentada uma descrição dos três principais modelos de serviço e dos modos de implantação de nuvens existentes. Para encerrar o capítulo, foi abordado o termo de Computação Verde e sua importância para a Computação em Nuvem, e também foram discutidos alguns dos principais

desafios relacionados aos projetos de infraestrutura dos *data centers* tradicionais, destacando principalmente o alto consumo de energia elétrica.

3 SOLUÇÃO PROPOSTA

Neste capítulo são apresentados, inicialmente, os conceitos e modelos referentes ao algoritmo de predição que constitui o objetivo principal deste trabalho, assim como a importância da abordagem que será proposta. Na seção 3.2, é descrito o modelo de previsão escolhido para a implementação do algoritmo. Na seção 3.3, a metodologia utilizada para a validação do algoritmo será mostrada. Em seguida, a seção 3.4 traz uma breve explicação acerca da implementação do algoritmo em um ambiente computacional em nuvem, bem como detalha o funcionamento do mesmo de acordo com as propriedades das infraestrutura de Computação em Nuvem.

3.1 Modelos de Previsão em Séries Temporais

Previsão de séries temporais é um desafio da área de Mineração de Dados. Prever valores futuros, em função de valores passados, tem se tornado um assunto de especial interesse na academia e na indústria, como por exemplo, em aplicações no mercado de ações, gastos mensais de energia de uma empresa, média mensal de clientes, etc. Vale ressaltar, que existem vários modelos que auxiliam na tarefa de previsão de séries temporais (RIBEIRO et al., 2009).

Uma série temporal, também denominada série histórica, é uma sequência de dados obtidos em intervalos regulares de tempo durante um período específico (MORETTIN; TOLOI, 1985). Este conjunto pode ser obtido através de observações periódicas do evento de interesse como, por exemplo, o total mensal de gasto energético em um data center. Se a série histórica for denominada como Z, o valor da série no momento t pode ser escrito como Zt (t = 1, 2, . . . , n) (Ver Figura 9) . Diz-se que uma série histórica é uma amostra de um processo estocástico (LATORRE; CARDOSO, 2001). Ou seja, dado um processo aleatório (não-determinístico) que varia no tempo, uma série temporal é um conjunto de observações discretas, realizadas em períodos equidistantes (SILVA et al., 2007) . Adicionalmente, uma série pode ser composta por três componentes não observáveis (MORETTIN; TOLOI, 1985): tendência, sazonalidade e a variação aleatória denominada de ruído branco. A tendência, é quando suas observações ocorrem aleatoriamente ao redor de uma média não-constante; define-se um fenômeno sazonal como aquele que ocorre regularmente em períodos fixos de tempo; e o ruído branco é conceituado como o tamanho da variação aleatória.

Segundo Morettin e Toloi (1985), os objetivos de se analisar uma série temporal são os seguintes:

- Descrição: propriedades da série como, por exemplo, o padrão de tendência, a existência de alterações estruturais, etc;

- Explicação: construir modelos que permitam explicar o comportamento da série no período observado;
- Controle de Processos: por exemplo, controle estatístico de qualidade;
- Previsão: prever valores futuros com base em valores passados.

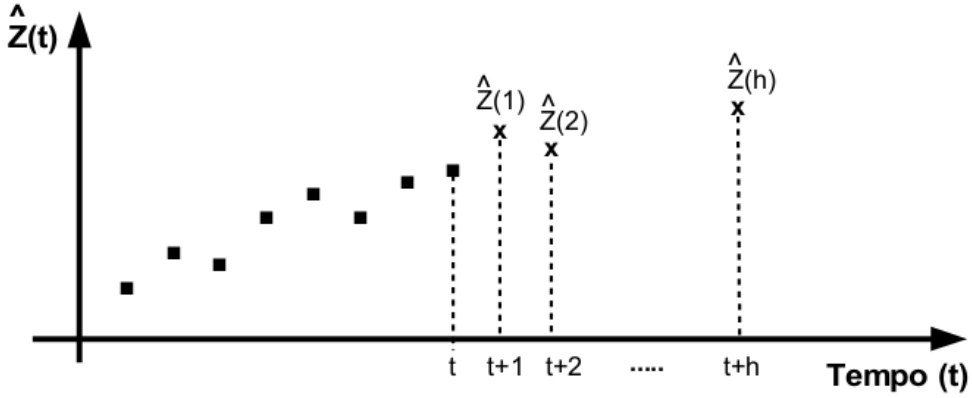

Figura 9: Observações de uma série temporal com previsões de origem t e horizontes de previsão iguais a um, dois e h. Extraída de MORETTIN e TOLOI (2006)

Como discutido no início da seção, existem vários modelos (ou técnicas) para previsão de séries temporais, dentre os quais, há os automáticos e os não-automáticos (MORETTIN; TOLOI, 1985):

- Automáticos: que são aplicados diretamente com a utilização de um computador;
- Não-automáticos: exige a intervenção de pessoal especializado, para serem aplicados.

Cabe destacar que, para a execução do presente trabalho, foi feita uma ponderação somente sobre os modelos automáticos, já que o algoritmo será executado por um sistema computacional e o uso de intervenções humanas é desnecessário.

Para exemplificar uma série real nesta seção, é utilizado um *Trace* disponibilizado pelo *Google* (REISS et al., 2012). Um *Trace* é um conjunto de arquivos formados por tabelas, cada qual com colunas que representam a quantidade de recursos requisitados. A geração dos *traces* foi feita por meio de um *cluster* do *Google* constituído por várias MFs. Então, todos os recursos requisitados são somente de máquinas reais.

Para este exemplo, foram utilizados os valores da coluna *Resource Request for CPU Cores* (coluna 10) contida na tabela *Task Events* do arquivo *part-00001-of-00500*, que totaliza cerca de 77.000 valores de *cores*. Estes valores representam a quantidade de *cores* de CPU requisitada. Vale ressaltar, que para a execução deste trabalho foram usados os mesmos valores.

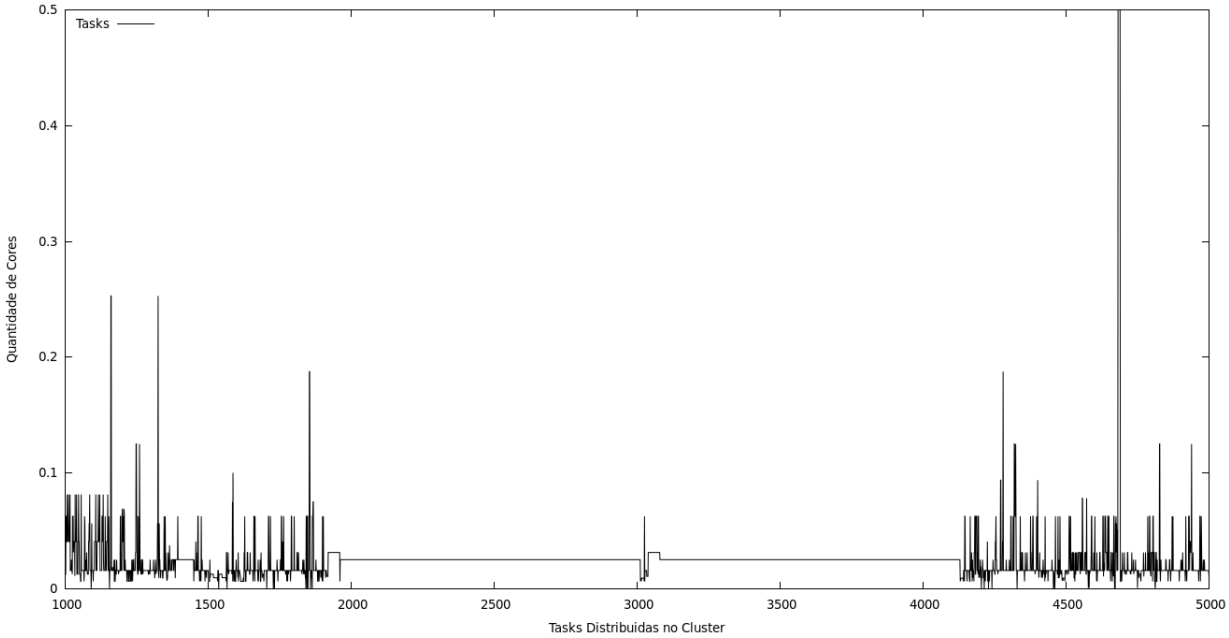

Figura 10: Carga de Trabalho dos *Clusters* do *Google*.

3.2 Alisamento Exponencial Simples

O modelo que será utilizado pelo algoritmo proposto é o de Alisamento Exponencial Simples. O motivo para isso, é que os métodos de alisamento possuem uma grande popularidade por não serem complexos, serem computacionalmente eficientes e obterem previsões razoáveis (MORETTIN; TOLOI, 1985). O AES pode ser descrito através da equação 3.1:

$$\overline{Z}_t = \alpha Z_t + (1-\alpha)\overline{Z}_{t-1}, \quad t = 1,...,N \tag{3.1}$$

onde $\overline{Z}$t é denominado o valor exponencialmente alisado, previsão para o período t, α é a constante de alisamento, $0 \leq \alpha \leq 1$, e Zt é o valor real (t-1).

Um detalhe importante sobre o AES, é que ele dá pesos maiores às observações mais recentes. Além disso, ele elimina a necessidade de armazenamento das observações anteriores, pois a formula necessita apenas da observação mais recente, a previsão imediatamente anterior e o valor de α (MORETTIN; TOLOI, 1985).

Sobre o valor de α, Ricardo (2009) afirma que não depende da escala em que as observações foram medidas, mas sim das características da série temporal. O valor de α deve ser especificado de modo a refletir a influência das observações passadas nas previsões. Valores pequenos produzem previsões que dependem de muitas observações passadas. Por outro lado, valores próximos de 1 levam a previsões que dependem das observações mais recentes e no caso extremo α = 1 a previsão é simplesmente a última observação.

Após a escolha do valor de α, o passo seguinte é determinar um valor de previsão inicial. Geralmente, segundo Delurgios (1998), o primeiro valor real é escolhido como a previsão para o segundo período. Por exemplo:

Período	Valor real	Valor previsto
1	700	
2	800	700

Figura 11: Escolha do valor de previsão inicial.

Sendo assim, para calcular a previsão para o período três, supõe-se que o valor inicial previsto para o período dois é igual ao valor real para o período um. Este método é muito eficiente e comum para o modelo AES.

Ressaltando os motivos da escolha deste modelo, segundo Morettin e Toloi (1985) o AES é amplamente utilizado devido às vantagens descritas abaixo:

- Fácil entendimento;
- Aplicação não dispendiosa;
- Grande flexibilidade permitida pela variação da constante de alisamento α;
- Necessidade de armazenar somente $\overline{Z}$t, Zt e α.

A principal desvantagem é a dificuldade em determinar o valor mais apropriado da constante de alisamento.

Considerando-se o fato de que determinação do valor de α é o principal desafio do AES, foi apresentado em Morettin e Toloi (1985) o modelo de Trigg e Leach, no qual a constante de alisamento varia automaticamente quando ocorre uma mudança no padrão básico da série. A definição de α é baseada em:

$$\alpha_t = \frac{E_t}{M_t}, \quad t = 1, ..., N \tag{3.2}$$

onde $E_t = \beta e_t + (1-\beta)E_{t-1}$, $M_t = \beta|e_t| + (1-\beta)M_{t-1}$, $\beta = 0,1$ ou 0,2. O erro de previsão no instante t é: $e_t = Z_t - \overline{Z}_{t-1}$. Adicionalmente, na primeira previsão α assume o valor de β.

Este método de atribuir valores dinâmicos à constante de alisamento, é baseado no seguinte argumento: aumentar a constante quando o sistema estiver fora de controle (S_t próximo de ± 1), dando pesos maiores aos valores mais recentes, e diminuir o valor de α quando o sistema estiver estiver sob controle ($S_t \cong 0$).

Uma desvantagem dessa abordagem, é que sua extensão para modelos que utilizam várias constantes, não é muito clara (MORETTIN; TOLOI, 1985).

3.3 Procedimentos Metodológicos

O primeiro passo para a execução deste trabalho, foi escolher qual modelo de predição utilizar. Foram escolhidos o AES e o modelo adaptativo de Trigg e Leach. Assim, possibilitando uma abordagem que permite que a constante de alisamento do modelo AES altere seu valor de acordo com o comportamento estatístico da série (ver Figura 10).

O segundo passo, consistiu na implementação do algoritmo, utilizando os modelos escolhidos no passo anterior. Nesse passo, optou-se por utilizar a linguagem de programação *Python*, por ser uma linguagem que oferece boa flexibilidade.

O terceiro passo, baseou-se na obtenção da quantidade de CPU requisita por cada *task* pertencente ao *trace* utilizado neste trabalho.

No quarto passo, foi considerada a lista de dados adquirida no passo anterior e dada como entrada para o algoritmo de predição implementado. Os valores foram observados a cada segundo. Nessa parte, o algoritmo utilizou o modelo AES juntamente com o modelo adaptativo de Trigg e Leach para prever os valores dos *cores*.

Por último, os dados obtidos pela previsão foram comparados com os dados da série real, no sentido de relatar a eficiência do algoritmo em obter valores futuros. Ou seja, o algoritmo é eficiente se for constatada a proximidade dos dados gerados pelo mesmo com os dados da série real.

3.4 Algoritmo de Predição

Como foi discutido no capítulo 2, os *data centers* tradicionais de Computação em Nuvem necessitam ser repensados com o intuito de atender aos requisitos da Computação Verde. Desta forma, os provedores de nuvem podem obter o mesmo desempenho, porém com a redução dos custos relacionados ao consumo energético. Beloglazov e Buyya (2010a) mostrou um modelo de arquitetura fortemente baseada na Computação Verde, que através da mesma é possível adicionar funcionalidades ao ambiente computacional envolvido, incrementando componentes que podem ser utilizados pelo sistema de gerenciamento, dando-o a possibilidade de gerenciar a migração e desligamento de MVs, consumo de energia, e tomar as demais decisões que o competem, de maneira consistente. Todavia, a arquitetura proposta não é definitiva, pois ainda há um amplo estudo para atingir esse objetivo.

O algoritmo que é a principal proposta do presente trabalho, pode ser visto como um

componente adicional para a arquitetura citada. Observando os elementos que a compõem, pode-se notar que o acréscimo de um componente que possa prever os valores de processamento das cargas de trabalho deve beneficiar no mínimo dois componentes: Gerenciador de MVs e o Monitor de Energia. Dado que o Monitor de Energia é o responsável pelo ligamento e desligamento das MVs, a execução desta tarefa pode ser ajudada pelo algoritmo de predição, pois saber um valor aproximado dos recursos que serão requisitados é uma vantagem em relação às tomadas de decisões. Já o Gerenciador de MVs, que é responsável pela migração de máquinas virtuais entre máquinas físicas, também pode obter os dados disponibilizados pelo componente de predição, facilitando a migração das VMs no sentido de liberar os recursos subutilizados de acordo com a demanda prevista dos usuários.

Vale ressaltar que o algoritmo é aplicado na quantidade de *cores* de processamento requisitada de cada carga proveniente dos usuários. Segundo Mishra et al. (2009), o uso de CPU destas cargas pode variar de 0-4 *cores*. Pode-se concluir, então, que é através da previsão da quantidade de processamento de cada carga que os componentes envolvidos poderão determinar se a quantidade de processamento alocado é suficiente à demanda dos usuários. De forma sucinta, a tomada de decisão do sistema de gerenciamento consiste em liberar ou alocar recursos, o que equivale a ligar ou desligar servidores de processamento.

Para a previsão das cargas de trabalho, o algoritmo emprega o modelo AES com o modelo adaptativo de Trigg e Leach, como já mencionado, e recebe como parâmetros duas listas de valores, valores reais e valores previstos, que são aplicados nos passos posteriores do mesmo.

3.5 Resumo

O capítulo 3 teve como principal objetivo introduzir os modelos de previsão de valores, bem como apresentar a metodologia utilizada e corroborar a importância do algoritmo proposto. De início, foi dada a definição de séries temporais e dos modelos de predição automáticos e não-automáticos. Em seguida, os objetivos de se analisar séries temporais foram apresentados, assim como a importância de se utilizar técnicas para perver dados futuros. Posteriormente, foi discutida a técnica de Alisamento Exponencial Simples, apresentando suas principais vantagens e desvantagens e os motivos pelo os quais o mesmo foi escolhido para execução deste trabalho. Em seguida, foi apresentada a metodologia empregada para execução da proposta. Por último, todavia não menos importante, foi explicado o algoritmo implementado e como ele pode contribuir com os sistemas de gerenciamento que buscam a Computação Verde.

4 RESULTADOS

Este capítulo apresenta e descreve os resultados obtidos deste trabalho. Especificamente, os resultados obtidos da execução do algoritmo proposto. Todos são mostrados em uma única seção, a qual aborda os resultados do algoritmo com variações na constante de alisamento.

4.1 Execução do algoritmo

O algoritmo foi desenvolvido totalmente baseado no modelo de Alisamento Exponencial Simples. Entretanto, observado o desafio de determinar o valor da constante de alisamento, o algoritmo utilizou também o modelo de Trigg e Leach para variar esse valor de acordo com o comportamento estatístico da série. Ele recebe como parâmetro a lista de valores de processamento obtida dos *traces* , observadas a cada segundo, outra lista contendo no mínimo o valor real e previsto da observação passada. No caso da primeira previsão, ou seja, t=3, assume-se que o valor previsto de t=2 é igual ao valor real de t=1.

Desta forma, para a execução deste trabalho, foram obtidos cerca de 77.000 dados que representam a quantidade de cores de processamento. Cada dado foi observado a cada segundo, totalizando em um tempo aproximado de 77.000 s de obervações de dados.

Os resultados podem ser vistos nos quatro gráficos abaixo:

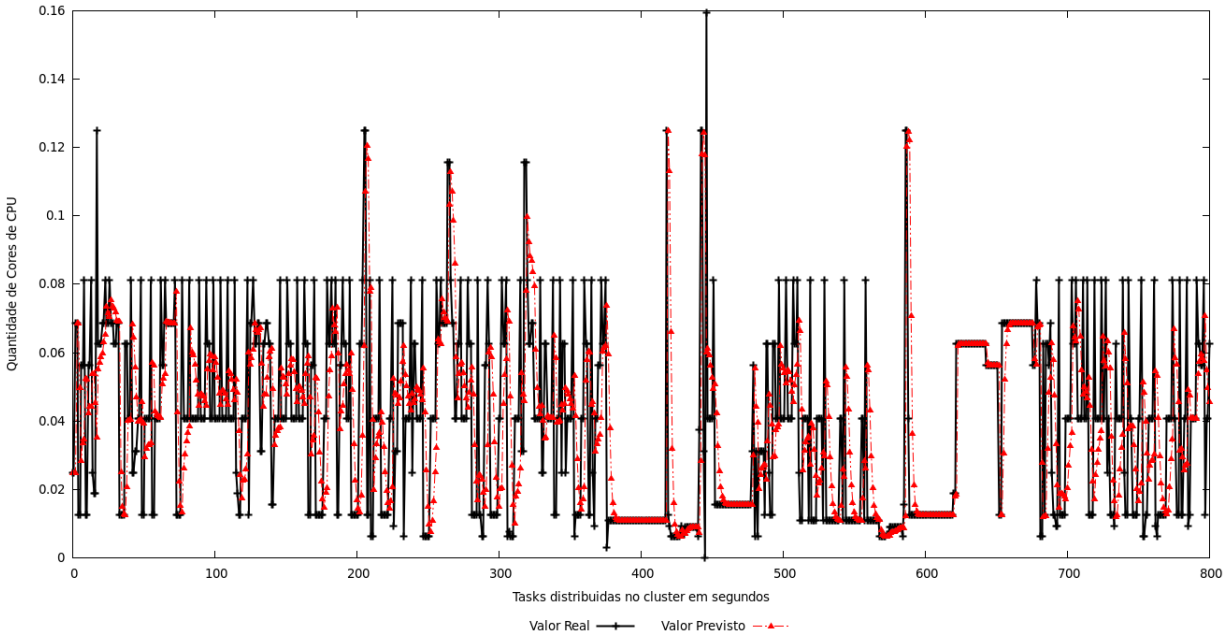

Figura 12: Previsões da quantidade de *cores* no intervalo de 0-800.

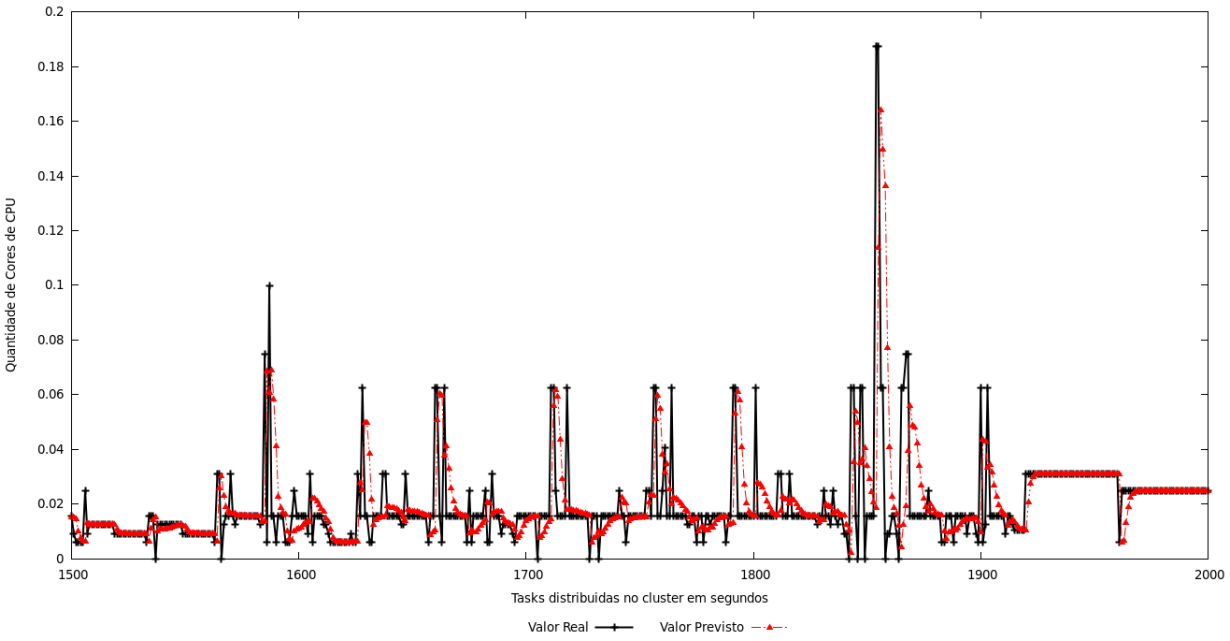

Figura 13: Previsões da quantidade de *cores* no intervalo de 1500-2000.

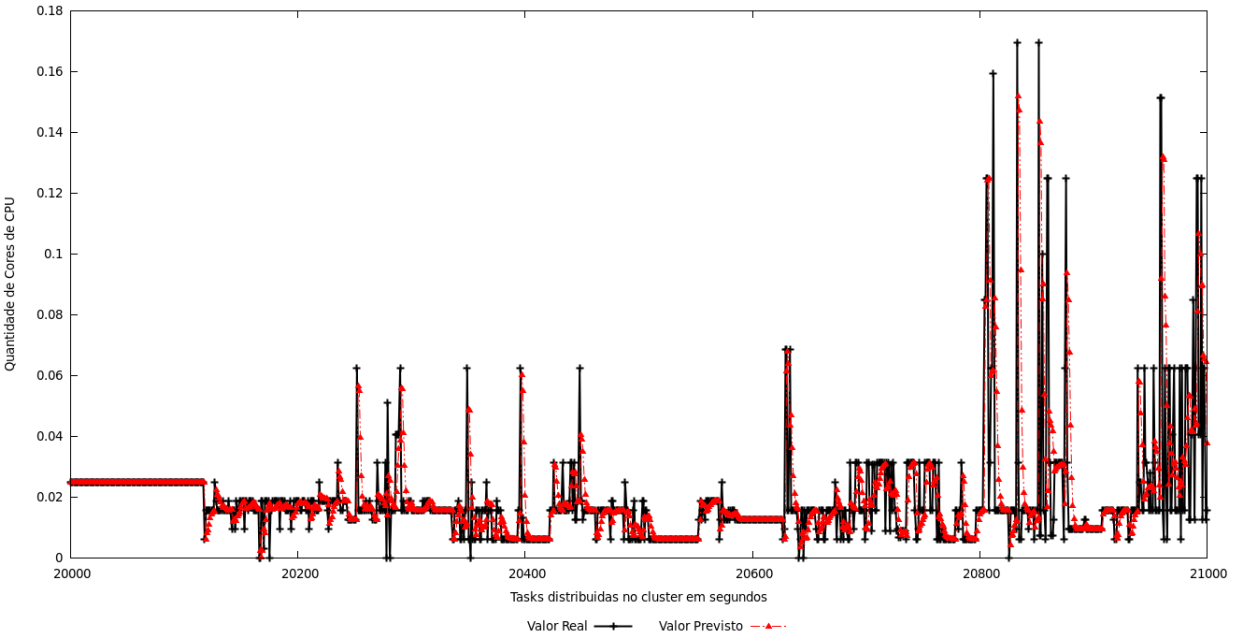

Figura 14: Previsões da quantidade de *cores* no intervalo de 20000-21000.

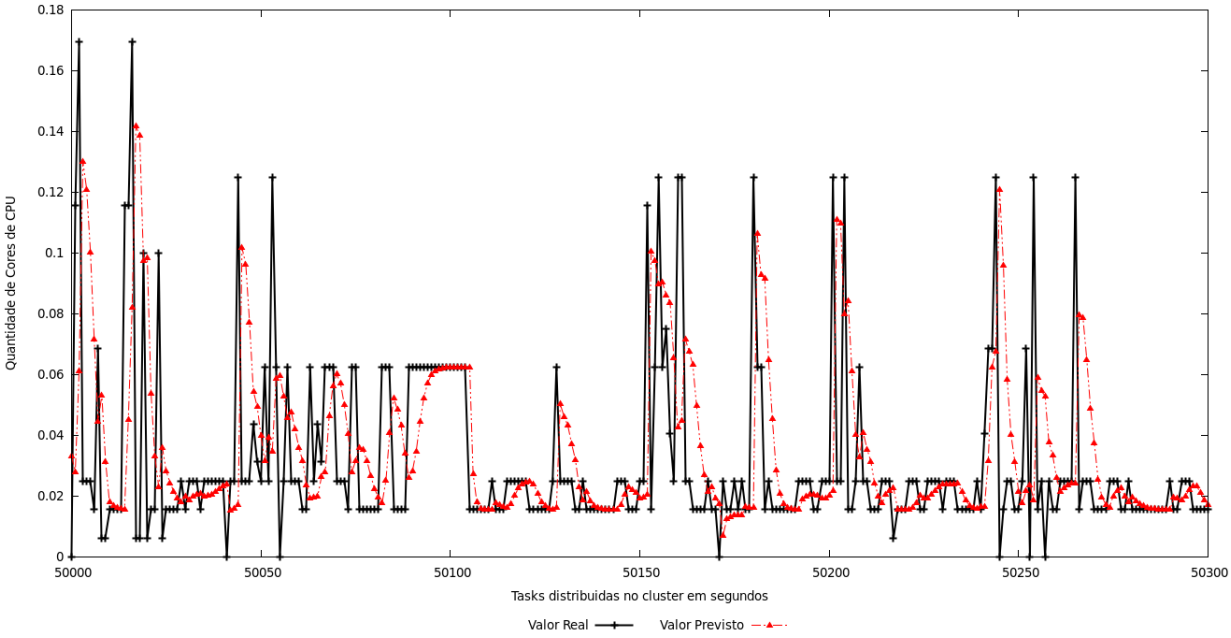

Figura 15: Previsões da quantidade de *cores* no intervalo de 50000-50300.

Como pode ser visto, o algoritmo se comportou de forma mais eficiente quando as variações dos valores de processamento reais não foram tão altas. Pois quando a série varia abruptamente nesse modelo, consequentemente são gerados maiores valores do erro de previsão (e_t) e de α, o que implica em um sistema fora de controle. Na Figura 14, com o intervalo entre 20000 e 21000, que, em certos pontos, os valores variaram de forma mais suave, pode-se notar uma maior aproximação dos valores. Em outras palavras, séries que estão sob controle (erros de previsão pequenos), ou seja, séries que não sofrem variações abruptas, produzem o valor de α próximo de 0.

Na Figura 12, com o inveralo entre 0 e 800, que conta com uma grande variação dos valores em certos momentos, nota-se que o algoritmo se mantém em uma média desses valores até que estes se encontrem em flutuações mais suaves.

A Figura 13 apresenta variações menores, em relação à Figura 12. Portanto, as previsões forão mais precisas.

Já na Figura 15, os resultados são mostrados em um intervalo menor, que pode-se notar de uma melhor forma como o algoritmo se adapta com os valores da série temporal.

4.2 Resumo

Este capítulo discutiu os resultados obtidos por meio da execução da proposta deste trabalho. Na seção 4.1 foram apresentados os gráficos dos resultados obtidos.

5 CONCLUSAO

Atualmente a Computação em Nuvem aparece como um novo paradigma em evolução, ela pode oferecer a seus usuários recursos computacionais sob demanda. Os recursos são oferecidos como serviços, que os usuários só pagam pelo o que usarem. Além disso, a *cloud computing* traz vários outros benefícios, como a redução dos custos na aquisição de *hardware* e *software* para os usuários finais. No entanto, os custos para implantação e manutenção dos *data centers* tradicionais de Computação em Nuvem têm preocupado os investidores. O alto consumo energético por parte das infraestruturas tem virado alvo de grandes esforços na tentativa de atenuar este problema, que consiste não somente em dinheiro gasto, mas também na poluição ambiental gerada.

A Computação Verde aparece nesse contexto como um modelo que objetiva mudanças nos projetos atuais de *data centers*. Especificamente, os ambientes computacionais em nuvem devem adotar novas técnicas que gerem melhor toda a heterogeneidade de hardware e software envolvida. Como por exemplo, a aquisição de sistemas de gerenciamento que possam tomar decisões importantes para a redução dos gastos. O desligamento de servidores subutilizados é um exemplo clássico disso.

Tendo como base os desafios expostos, o presente trabalho teve como proposta a implementação de um algoritmo que fosse capaz de prever a quantidade de *cores* de processamento requisitada pelas cargas de trabalho, com o objetivo de prover informações adicionais ao sistema de gerência de forma a facilitar as tomadas de decisões do mesmo.

Os resultados obtidos com a execução do algoritmo mostraram-se satisfatórios quando não houve grandes flutuações das cargas reais. Nos momentos em que as cargas mantiveram-se com flutuações suaves, as previsões mostraram-se mais precisas, com o algoritmo conseguindo prever bem as variações das cargas de trabalho. Apesar de o algoritmo não dispor da mesma precisão quando a série sofre variações abruptas, acredita-se que os resultados obtidos aproximem-se estatisticamente do padrão da série, o que pode facilitar ao sistema de gerência identificar um valor médio dos recursos computacionais requisitados. Desta forma, a proposta deste trabalho pode incrementar as funcionalidades dos sistemas de gerenciamento de recursos que adotem o uso da computação verde, servindo como um componente adicional na infraestrutura.

Com esses resultados, espera-se ter contribuido para o campo de pesquisa e desenvolvimento na área da Computação em Nuvem.

5.1 Perspectivas para trabalhos futuros

Como trabalho futuro, será desenvolvido um ambiente computacional em nuvem por meio da ferramenta *OpenNebula* (OPENNEBULA, 2013), e estudado um modelo de predição que preveja as cargas de trabalho com maior precisão. O objetivo é testar um ambiente real. O uso da ferramenta *httperf* (HTTPERF, 2013) para a geração das cargas de trabalho é considerado uma possibilidade.

BIBLIOGRAFIA

ANDREAS, B. et al. Energy-efficient cloud computing. In: *The Computer..Vol 53*. ENG: Oxford University Press on behalf of The British Computer Society, 2009.

BADGER, L. et al. *Recommendations of the National Institute of Standards and Technology*. U.S. Department of Commerce. Computer Security Division Information Technology, Laboratory National Institute of Standards and Technology,Gaithersburg, Maio 2011.

BARROSO, L. A.; HöLZLE, U. The case for energy-proportional computing. In: *Computer*. USA: IEEE Computer Society, 2007. p. 33–37.

BELOGLAZOV, A.; BUYYA, R. Energy-efficient management of data center resources for cloud computing: A vision, architectural elements, and open challenges. In: *International Conference on Parallel and Distributed Processing Techniques and Applications*. Las Vegas, Usa: (PDPTA 2010), 2010.

BELOGLAZOV, A.; BUYYA, R. Energy efficient resource management in virtualized cloud data centers. In: *10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*. Melbourne, Australia: IEEE Computer Society, 2010.

BINDER, W.; SURI, N. Green computing: Energy consumption optimized service hosting. In: *SOFSEM '09: Proceedings of the 35 Conference on Current Trends in Theory and Practice of Computer Science*. Spindleruv Mlýn, Czech Republic: Springer, 2009. p. 117–128.

BRANTNER, M. et al. Building a database on s3. In: *2008 ACM SIGMOD international conference on Management of data - SIGMOD'08*. New York: ACM Press, 2008. p. 251.

CARISSIMI, A. Virtualização: da teoria a soluções. In: *26 Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. Porto Alegre, Brasil: UFRGS, 2008. p. 174–207.

CARISSIMI, A. Virtual putty: Reshaping the physical footprint of virtual machines. In: *Workshop on Hot Topics in Cloud Computing*. US: HotCloud'09, 2009.

COULOURIS, G.; DOLLMORE; KINDBERG. *Distributed Systems: Concepts and Design*. [S.l.]: bookman, 2005.

DELURGIOS, S. A. *Forecasting principles and applications*. Singapura: McGraw-Hill, 1998.

GREENBERG, A. et al. Vl2: A scalable and flexible data center network. In: *ACM SIGCOMM 2009 Conference*. Barcelona, Spain: http://blog.rightscale.com/2008/05/26/define-cloud-computing, 2009.

HAMILTON, J. Diseconomies of scale. Blog post. 2008.

HTTPERF. http://www.hpl.hp.com/research/linux/httperf/. Acesso em Julho 2013. 2013.

LAGO, D. G.; MADEIRA, E. R. M.; BITTENCOURT, L. F. Power-aware virtual machine scheduling on clouds using active cooling control and dvfs. In: *9th International Workshop on Middleware for Grids, Clouds and e-Science*. Lisbon, Portugal: ACM 2011 Article, 2011.

LAGO, D. G.; MADEIRA, E. R. M.; BITTENCOURT, L. F. Escalonamento com prioridade na alocação ciente de energia de máquinas virtuais em nuvens. In: *XXX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. Minas Gerais, Brasil: UFMG, 2012.

LATORRE, M. D. de O.; CARDOSO, M. R. A. Análise de séries temporais em epidemiologia: uma introdução sobre os aspectos metodológicos. In: *CNPq*. São Paulo, Brasil: [s.n.], 2001.

MAHOWALD, R. P. Market analysis perspective: Worldwide saas and cloud services 2010: New models for delivering software. In: *IDC*. International Data Corporation: International Data Corporation, 2011.

MISHRA, A. K. et al. Towards characterizing cloud backend workloads:insights from google compute clusters. In: *Interning Google*. Mountain View, USA: Google, 2009.

MORETTIN, P. A.; TOLOI, C.M.C. *Previsão de séries temporais*. São Paulo, Brasil: Atual Editora, 1985.

MORETTIN, P. A.; TOLOI, C. M. C. *Análise de Séries Temporais*. São Paulo, Brasil: Edgard Blücher, 2006.

OPENNEBULA. http://www.opennebula.org. Acesso em julho de 2013. 2013.

PALLIS, G. Cloud computing: The new frontier of internet computing. view from the cloud. In: *Cloud'2010*. US: Comp., 2010.

QURESHI, A. et al. Cutting the electric bill for internet-scale systems. In: *SIGCOMM '09: Proceedings of the ACM SIGCOMM 2009 conference on Data communication*. Barcelona, Spain: ACM 2009 Article, 2009.

REISS, C.; WILKES, J.; HELLERSTEIN, J. Google cluster-usage traces: format + schema. In: *Google Inc.* USA: Google, 2012.

RIBEIRO, C. V.; GOLDSCHMIDT, R. R.; CHOREN, R. *Métodos para Previsão de Séries Temporais e suas Tendências de Desenvolvimento*. Dissertação (Mestrado) — INSTITUTO MILITAR DE ENGENHARIA, Rio de Janeiro, Brasil, 2009.

RICARDO, S. E. *Análise de Séries Temporais*. Brasil: RICARDO SANDES EHLERS, 2009.

RIGHTSCALE. Define cloud computing. In: *RightScale Blog*. US: http://blog.rightscale.com/2008/05/26/define-cloud-computing, 2010.

SHULZ, G. *The Green and Virtual Data Center*. [S.l.]: CRC Press - Taylor Francis Group, 2004.

SILVA, P. O. M. P. et al. Previsão de séries temporais utilizando lógica nebulosa. In: *4o CONTECSI*. São Paulo, Brasil: USP, 2007.

SOUSA, F. R. C.; MOREIRA, L. O.; MACHADO, J. C. Computação em nuvem: Conceitos, tecnologias, aplicações e desafios. In: EDUFPI (Ed.). *ERCEMAPI*. Piauí, Brasil: UFPI, 2009.

STANOEVSKA-SLABEVA, K. Grid and cloud computing - a business perspective on technology and application. In: *Cloud'2010*. US: Springer, 2010.

SUN. Define cloud computing. In: *Sun whitepaper*. US: https://www.sun.com/offers/docs/cloudcomputingprimer:pdf:ltimoacessoem10=10=2010, 2010.

TANENBAUM, M. V. S. A. S. *Distributed Systems. Principles and Paradigms*. [S.l.]: Addison-Wesley, 2006.

VAQUERO, L. M. et al. Clouds: Towards a cloud definition. In: *SIGCOM*. Comunn: SIGCOM, 2009. p. 50–55.

VERDI, F. L. et al. Novas arquiteturas de data center para cloud computing. In: *Simposio Brasileiro de Redes de Computadores*. Brasil: SBRC, 2010.

VMWARE. The benefits of virtualization for small and medium businesses. Acesso em julho de 2013. 2013.

WERNER, Jorge. *UMA ABORDAGEM PARA ALOCAÇÃO DE MÁQUINAS VIRTUAIS EM AMBIENTES DE COMPUTAÇÃO EM NUVEM VERDE*. Dissertação (Mestrado) — UNIVERSIDADE FEDERAL DE SANTA CATARINA - CENTRO TECNOLÓGICO, Florianólopis, Brasil, 2011.

WERNER, J. et al. Aperfeiçoando a gerência de recursos para nuvens verdes. In: *Universidade Federal de Santa Catarina*. Florianópolis – SC – Brasil: UFSC, 2012.