UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Otimização da Alocação de Recursos em Data Centers Hierarquicamente Distribuídos

Rafael Fogarolli Vieira

DM 41/2018

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2018

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Rafael Fogarolli Vieira

Otimização da Alocação de Recursos em Data Centers Hierarquicamente Distribuídos

DM 41/2018

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2018

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Rafael Fogarolli Vieira

Otimização da Alocação de Recursos em Data Centers Hierarquicamente Distribuídos

Dissertação submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica da UFPA para a obtenção do Grau de Mestre em Engenharia Elétrica na área de Computação Aplicada.

Orientador: Diego Lisboa Cardoso

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2018

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

"OTIMIZAÇÃO DA ALOCAÇÃO DE RECURSOS EM DATA CENTERS HIERARQUICAMENTE DISTRIBUÍDOS"

AUTOR: **RAFAEL FOGAROLLI VIEIRA**

DISSERTAÇÃO DE MESTRADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM: 19/12/2018

BANCA EXAMINADORA:

Diego Lisboa Cardoso

Prof. Dr. Diego Lisboa Cardoso

(Orientador – PPGEE/UFPA)

Eduardo Cerqueira

Prof. Dr. Eduardo Coelho Cerqueira

(Avaliador Interno – PPGEE/UFPA)

Carlos N. da Silva

Prof. Dr. Carlos Natalino da Silva

(Avaliador Externo – KTH/SUÉCIA)

VISTO:

Prof.ª Dr.ª Maria Emília de Lima Tostes

(Coordenadora do PPGEE/ITEC/UFPA)

Agradecimentos

Primeiramente, agradeço a Deus, por sua misericórdia e amor infinito.

Aos meus pais, Júlio e Cláudia, pelo apoio incondicional, sacrifícios feitos, incentivo e compreensão, mesmo nas horas mais difíceis.

À Universidade Federal do Pará pela oportunidade de fazer a Pós Graduação, em especial ao PPGEE, e a todos os profissionais envolvidos.

Agradeço a todos os professores do Programa que tive a oportunidade de conhecer, que dividiram comigo conhecimento e contribuíram para a minha formação.

Ao meu orientador, professor Diego Cardoso, pela orientação, amizade e confiança depositada em mim ao longo do tempo. A todos os meus amigos (a) do LPO, Daniel, Ermínio, Carlos, Paulo Henrique, Welton, Paulo, Matheus Leto, Rita, Mariane, Igor, Nilze e aos demais que compartilharam comigo a experiência única que é fazer parte de um Programa de Pós-Graduação.

Aos amigos do LCT, LEA, LINC, e LPRAD, que fizeram parte dessa aventura insana e que compartilharam momentos únicos ao longo dessa trajetória.

Agradeço ao professor Carlos Natalino, que mesmo não estando presente, dedicou seu tempo para me orientar e me incentivar a continuar no caminho acadêmico.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo suporte financeiro ao longo do programa.

"Tudo o que temos de decidir é o que fazer com o tempo que nos é dado".

(Gandalf)

Resumo

O crescente volume de serviços e aplicativos, além do crescimento acelerado das demandas de acesso sem fio, representam desafios significativos para a próxima geração de redes móveis. Esse aumento no volume de aplicações é reflexo do crescimento de "coisas"que estão se conectando na rede e irão produzir e consumir dados de maneira exorbitante. Um novo paradigma que está ganhando reconhecimento no contexto de redes sem fio, responsável por parte desse crescimento elevado do volume de serviços e aplicações, é a Internet das Coisas. Conectar esses dispositivos irá leva uma enorme quantidade de informação a trafegar na rede que irá exigir recursos significativos de computação para serem processados e armazenados. Uma abordagem proeminente para tratar esses dados é o uso da computação em nuvem, a qual utiliza *datacenters* para armazenamento e processamento dos dados. Entretanto, o aumento exponencial de dados gerados e de aplicações com requisitos mais severos, tornam a nuvem tradicional, que concentra seus recursos no núcleo da rede, menos adequada para cenários envolvendo estas operações, que exigem baixa latência e alta qualidade de serviço. Assim, para lidar com tais adversidades, um novo conceito emergente, conhecido como *Edge Cloud Computing*, foi proposto para se tornar uma extensão da computação em nuvem tradicional, aproximando recursos computacionais, escaláveis, para a borda da rede, criando assim, uma hierarquia de *datacenters*. Desta forma, os serviços e aplicações que possuem requisitos mais estritos conseguirão ter suas necessidades atendidas, como por exemplo, a obtenção de experiência de usuário quase instantânea. Sendo assim, este trabalho propõe formulação matemática para o dimensionamento e provisionamento de uma hierarquia de DCs. Através dos resultados obtidos, observa-se que a hierarquia de DCs atrelada ao modelo proposto consegue ser superior aos demais cenários, conseguindo alocar 99% do conjunto de aplicações que foram utilizado nos testes, e diminuir o fluxo nos *links* de *backhual*, o qual é gerado pelo maçante número de aplicações que iriam circular pela rede. As análises confirmam a necessidade da aproximação de recursos computacionais para a borda da rede, atrelada com uma eficiente estratégia de alocação de aplicações, a fim de garantir um melhor desempenho para a rede.

Palavras-chave: *Edge Cloud Computing*, computação em nuvem, *datacenters*, alocação de recursos.

Abstract

The rapidly increasing volume of services and applications, in addition to the high wireless access demand, are significant challenges for the next generation of mobile networks. The growth in the volume of applications is the reflection of the quantity of "things" that are being connected to the network and generating a huge data traffic. A new paradigm that is gaining recognition in the field of wireless networks, and is also responsible for part of this growth in the volume of services and applications, is the Internet of Things. The high amount of data that is generated by connecting those devices to the network will require significant computational resources to be processed and stored. A prominent approach to handling such large amount of data is the use of Cloud Computing, which uses datacenters for storage and data processing. However, traditional Cloud Computing, which has centralized resources, is not able to handle the high volume of data and the strict latency and Quality of Service requirements. Thus, to address such adversities, a new emerging concept known as Edge Cloud Computing has been proposed as an extension of the traditional Cloud Computing, bringing computational resources to the edge of the network and thereby creating a hierarchy of datacenters. In this way, the stricter requirements from services and applications, such as obtaining near-instant user experience, can be satisfied. In this work, a mathematical formulation for the dimensioning and provisioning of a hierarchy of DCs is proposed. According to the obtained results, the hierarchy of DCs provisioned and dimensioned using the proposed model can be better when compared to the others, being able to allocate 99% of the set of applications that were used in the tests and to decrease the data flow in the backhaul links that is generated by the high number of applications the would circulate through the network. The analysis highlight the necessity of bringing computational resources to the network's edge in addition to an efficient applications allocation strategy in order to guarantee a better network performance.

Keywords: Edge Cloud Computing, Cloud Computing, datacenters, resource allocation.

Lista de ilustrações

Figura 1 Dados móveis e tráfego da internet, 2016-2021. 1

Figura 2 Arquitetura da Computação em Nuvem 6
Figura 3 Modelos de Serviço em Nuvem. 8
Figura 4 Diferentes Geração da comunicação móvel 13
Figura 5 Áreas de aplicativos IoT. 15
Figura 6 Aplicações IoT 16

Figura 7 Rede Hierarquicamente Distribuída em Três Níveis. 23

Figura 8 Topologias de DCs 30
Figura 9 Utilização dos Links (%) 31
Figura 10 Uso Individual dos Links (%) 32
Figura 11 Utilização dos Recursos Computacionais (%) 33
Figura 12 Utilização dos Recursos Computacionais na Hierarquia de DCs (%) . . 34
Figura 13 Medias das aplicações alocadas (%) 35
Figura 14 Aplicações Alocadas por níveis de DCs 36
Figura 15 Caracterização das aplicações 36

Lista de tabelas

Tabela 1 Perfis de Aplicação 25

Tabela 2 Parâmetros das Topologias 29
Tabela 3 Parâmetros dos DC 30

Lista de abreviaturas e siglas

ACC	Acceleration Sensor
APCA	Adaptive Piecewise Constant Approximation
API	Application Programming Interface
AANN	Auto-Associative Neural Network
CHEB	Chebyshev Approximation
CSS	Composite Strain Sensor
CAGR	Compound Annual Growth Rate
CRUD	Create, Read, Update and Delete
DAO	Data Access Object
DAQ	Data acquisition
FBG	Fiber Bragg gratings
FCL	Fuzzy Control Language
GPTSO	General Purpose Temperature Sensor Outdoor
IEC	International Electrotechnical Commission
InterAB	Interrogator Abstraction
JDBC	Java Database Connectivity
MSD	Mahalanobis Squared Distance
MVC	Model-View-Controller
MIF	Motor de inferências Fuzzy
MGD	Multivariate Gaussian Distribution
NPCA	Nonlinear Principal Component Analysis
UGD	Univariate Gaussian Distribution

ORM	Object Relational Mapping
OIDA	Optoelectronics Industry Development Association
PCA	Piecewise Constant Approximation
PCA*	Principal Component Analysis
PWLH	PieceWise Linear Histogram
RMSE	Root Mean Square Error
SM	Single Mode
SF	Slide Filters
SVD	Singular Value Decomposition
SGBD	Sistema Gerenciador de Banco de Dados
SRF	Sistemas baseados em regras Fuzzy
SHM	Structural Health Monitoring
SQL	Structured Query Language
UML	Unified Modeling Language
WbS	Web-based System
WSS	Weldable Strain Sensor
WTS	Weldable Temperature Sensor
WSN	Wireless Sensor Networks
XHTML	EXtensible HyperText Markup Language
XML	EXtensible Markup Language

Lista de abreviaturas e siglas

1G Primeira Geração

2G Segunda Geração

3G Terceira Geração

4G Quarta Geração

5G Quinta Geração

AI Automação Industrial

AMPS *Advenced Mobile Phone System*

ASP *Application Service Provider*

BD *Backup* de Dados

BS *Base Station*

C-RAN *Centralized*-RAN

CI Casa Inteligente

CPU *Central Process Unit*

D-AMPS Digital-AMPS

DC *Data Center*

DCB *Datacenters* de Borda

DCN *Datacenters* de Núcleo

DCR *Datacenters* Regionais

ECC *Edge Cloud Computing*

ETSI *European Telecommunications Standards Institute*

FF *First-Fit*

FO Função Objetivo

FOG *Fog Computing*

Gbps Gigabit por segundo

GSM *Global System for Mobile communication*

HSPA *High Speed Packet Access*

Iaas *Infrastructure as a Service*

ILP *Integer Linear Programming*

IoT *Internet of Things*

IT Internet Tátil

LTE Long Term Evolution

M Médica

MA Monitoramento Ambiental

MAN *Metropolitan Area Network*

MatLab *MatrixLaboratory*

Mbps Megabit por segundo

MEC *Mobile Edge Computing*

ms milissegundo

NFV *Network Functions Virtualization*

PaaS *Plataform as a Service*

PDC *Personal Digital Cellular*

PL Programação Linear

PL Programação Linear

PO Pesquisa Operacional

QoS Qualidade de Serviço

RAN *Radio Access Network*

RV Realidade Virtual

SaaS *Software as a Service*

SG *Smart Grid* (SG)

SLA *Service Level Agreement*

T1 Topologia 1

T2 Topologia 2

TACS *Total Access Communication System*

TI Tecnologia da Informação

UP Unidade de Processamento

US Unidade de Armazenamento

v-FAP *Fog Access Point*

VM *Virtual Machine*

VMP *Virtual Machine Placement*

Sumário

1 Introdução **1**

1.1 Motivações . . . 3
1.2 Objetivos . . . 4
1.3 Organização da Dissertação . . . 5

2 Revisão Literária **6**

2.1 Considerações iniciais . . . 6
2.2 Computação em Nuvem . . . 6
2.2.1 Características Essenciais . . . 7
2.2.2 Modelos de Serviço . . . 8
2.2.3 Modelos de implantação . . . 9
2.3 Computação de Borda . . . 10
2.4 Otimização Linear . . . 11
2.5 Redes Móveis de Quinta Geração (5G) . . . 13
2.6 Aplicações - 5G . . . 14
2.7 Considerações finais . . . 16

3 Trabalhos Correlatos **17**

3.1 Considerações iniciais . . . 17
3.2 Alocação de Recursos . . . 17
3.3 Computação de Borda . . . 19
3.4 Hierarquia de Data Centers . . . 21
3.5 Considerações finais . . . 22

4 Formulação Matemática **23**

4.1 Considerações iniciais . . . 23
4.2 Topologia e Variáveis . . . 23
4.3 Função Objetivo . . . 25
4.4 Restrições . . . 26
4.5 Considerações finais . . . 28

5 Estudos de Caso e Resultados **29**

5.1 Considerações iniciais . . . 29
5.2 Estudo de Caso . . . 29
5.3 Resultados . . . 31
5.3.1 Utilização de Recursos . . . 31

5.3.2 Alocação das Aplicações 34
5.4 Considerações finais 36

6 Conclusões **38**
6.1 Contribuições da Dissertação 39
6.2 Contribuições Adicionais 39
6.3 Trabalhos Futuros 40
6.4 Dificuldades Encontradas 41

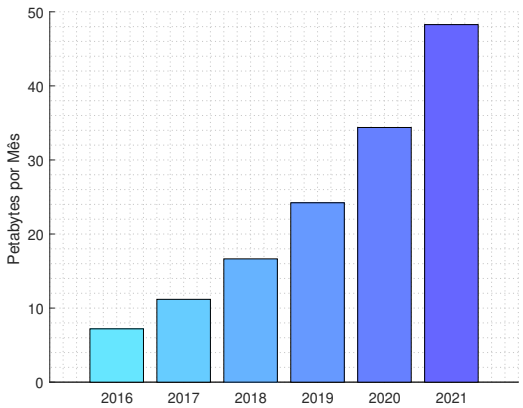
Referências **42**

1 Introdução

O crescente volume de serviços e aplicativos, além do crescimento acelerado das demandas de acesso sem fio, representam desafios significativos para a próxima geração de redes móveis, a quinta geração (5G). Globalmente, o tráfego de dados móveis aumentará sete vezes até 2021, crescendo a uma taxa de 46% entre os anos de 2016 e 2021, chegando a 8,3 *Exabytes* por mês até 2021, como é mostrado na Figura 1 (CISCO, 2017). Esse aumento no volume de serviços é reflexo do crescimento de "coisas"que estão se conectando na rede, as quais irão produzir e consumir dados de maneira exorbitante. Muitas dessas "coisas" farão parte de sistemas complexos, que exigirão capacidade de computação e armazenamento para processar e armazenar seus dados.

Um novo paradigma que está ganhando reconhecimento no contexto de redes sem fio e representa parte desse crescimento elevado do volume de serviços e aplicações é a Internet das Coisas (IoT - *Internet of Things*). A ideia central deste paradigma é usar todos os dispositivos que nos rodeiam, por exemplo, sensores, atuadores, telefones celulares, entre outros, a fim de cooperar para alcançar objetivos comuns (CAMPEANU, 2018). Conectar esses dispositivos irá levar uma enorme quantidade de informação a trafegar na rede, por exemplo, vídeos capturados por *smartphones* ou *tablets*, dados de sensores e atuadores, aplicações 3D e de realidade virtual, entre outras, que precisam ser armazenadas e/ou processados com eficiência, gerando uma sobrecarga de dados a serem trafegados na infraestrutura de rede. Desta forma, a IoT exigirá recursos significativos de computação, mas a questão de onde esses recursos devem ser colocados permanece aberta.

Figura 1 – Dados móveis e tráfego da internet, 2016-2021.



Adaptação de (CISCO, 2017)

Uma abordagem proeminente é o uso da computação em nuvem, a qual utiliza *data centers* para armazenamento e processamento de dados, e com isso, remove a necessidade de alta complexidade no *hardware* dos clientes. A computação em nuvem tradicional localiza os seus recursos computacionais no núcleo da rede, conhecidos como Data Centers (DC), que são capazes de gerenciar, armazenar e processar grandes volumes de dados.

O paradigma de *Data center* foi projetado para ser escalável, tanto em termos do poder de processamento quanto recursos de armazenamento. Tal característica torna-se, extremamente, importante devido a alta demanda do tráfego de dados e do poder de processamento e armazenamento advindas da ampla variedade de aplicativos e dispositivos IoT. Entretanto, o aumento exponencial de dados gerados e de aplicações com requisitos mais severos, tornam a nuvem tradicional menos adequada para cenários envolvendo estas operações, que exigem baixa latência e alta qualidade de serviço (QoS). De fato, decidir a localização de tais recursos coloca questões não triviais aos projetistas de plataformas.

Do ponto de vista do cliente, a localização desses recursos computacionais atribuídos a um serviço não é importante, desde que as suas restrições de QoS como taxa de dados, latência e disponibilidade, por exemplo, sejam atendidos. O uso da computação em nuvem pode causar um atraso significativo no tempo de resposta das aplicações e, estas, podem provocar uma sobrecarga significativa de tráfego no núcleo da rede (CORCORAN; DATTA, 2016) se não forem bem provisionadas. Outros fatores como o custo para transportar toda a quantidade de dados gerada na borda até núcleo da rede também é indesejável.

Para lidar com tais adversidades, um novo conceito emergente, conhecido como *Edge Cloud Computing* (ECC), foi proposto. A ideia do ECC é levar os serviços em nuvem ou parte deles para a rede de borda, que está localizada na proximidade geográfica dos usuários (WANG et al., 2017), desse modo, passa a existir escalabilidade de recursos computacionais em vários níveis da rede. Propostas baseadas em ECC estão surgindo para atender essas necessidades que são resultado da jovem geração de aplicações e serviços. Dessa forma, a ECC se torna uma extensão da computação em nuvem tradicional, aproximando recursos computacionais, escaláveis, para a borda da rede.

Uma hierarquia de DCs, então, começa a ser formada para atender as novas demandas de requisitos, onde os DCs mais próximos da borda da rede possuem menos recursos computacionais, entretanto as latências para alcançar o núcleo da rede podem ser um fator limitante para algumas aplicações que necessitam de execução em tempo real. Orquestrados e gerenciados de maneira inteligente, os recursos de computação e armazenamento colocados na borda da rede, podem lidar com o crescente número de dispositivos conectados e a demanda emergente em IoT.

Essa hierarquia de DCs é implementada com a intenção de facilitar o rápido desenvolvimento de sistemas complexos de computadores e atender aos seguintes requisitos

(SKALA et al., 2015):

- Desempenho: otimizado para respostas rápidas, alto processamento e baixa latência;
- Disponibilidade: requer redundância, recuperação rápida no caso de falhas de sistema;
- Confiabilidade: o sistema precisa ser confiável em dados e funções;
- Capacidade de Gerenciamento: sistema escalável com facilidade de operação.

1.1 Motivações

Atualmente, um dos grandes desafios em redes móveis é a necessidade de oferecer a melhor experiência possível para uma infinidade de aplicação e serviços, que estão surgindo junto com o 5G, e que possuem diferentes perfis e requisitos de QoS. A utilização de uma hierarquia de DCs, para fornecer a melhor experiência para esses usuários, se torna uma solução atrativa. As demandas por recursos de computação e armazenamento provavelmente virão de dezenas de bilhões de dispositivos fixos e móveis, que abrangerão vastas áreas geográficas e serão organizados de várias formas diferentes, cobrindo uma infinidade de casos de uso e configurações. Por sua vez, muitas dessas configurações terão requisitos rigorosos, como latência muito baixa, alta taxa de transferência durante períodos de tempo curtos, tomada de decisão rápida com base em análises em tempo real e combinações diferentes desses e de outros requisitos.

A partir do levantamento do estado da arte, descrito nos próximos capítulos, observa-se que a aproximação dos DCs para a borda da rede é uma temática atual e recorrente na comunidade acadêmico-científica, pois essa aproximação ocasiona a formação de um hierarquia de DCs, onde a ordenada distribuição de poderes de processamento e armazenamento decrescem com a aproximação dos DCs à borda da rede. A hierarquia de DCs que passa a existir oferece, então, computação escalável em vários níveis da rede, podendo atender esses diferentes perfis de aplicação e serviços. As aplicações poderão ser processadas em qualquer nível da rede, onde tenha um DC capaz de atende-la, mantendo as características da nuvem tradicional, porém mais próximos da borda, trazendo um conjunto de vantagens, como diminuição do tráfego no núcleo da rede e tempo de resposta muito menor que o atual.

Os serviços e aplicações que possuem requisitos mais estritos conseguirão ter suas necessidades atendidas, como exemplo, a obtenção de experiência de usuário, quase instantânea, exigida por muitas aplicações de internet tátil interativa que podem exigir latências, ponta-a-ponta, abaixo de 10 milissegundos. A aproximação dos DCs para a borda da rede está ganhando muita atenção, nos últimos anos, devido as suas vantagens em termos de

economia de custos e melhoria da experiência para as novas aplicações. Assim, a hierarquia de DCs pretende ter um papel fundamental na redução da latência em aplicativos dinâmicos que necessitam ser processados em tempo real.

A hierarquia de DCs surge para solucionar diversos problemas, como a redução da carga na nuvem centralizada, criando gargalos e pontos de falha originados pelo descomedido crescimento de serviços e aplicações da nova geração de redes móveis, e a redução da latência fim-a-fim, entre outros. Entretanto, a hierarquia de DCs deixa em aberto novos desafios para a academia e a indústria, como o dimensionamento e provisionamento inteligente dos novos recursos presentes na borda da rede e a segurança dos dados que estão transitando por ela.

1.2 Objetivos

Nesta dissertação, uma formulação matemática para o provisionamento de uma hierarquia de DCs é proposta. Essa formulação comporta não só as avaliações realizadas no estudo de caso, mas também pode ser usada para contribuir em diversos outros problemas que necessitem ser modelados e resolvidos utilizando programação linear (PL). É proposto também, a utilização do *First-Fit* (FF) como *benchmark* para comparação com os resultados obtidos no modelo. O algoritmo FF é o mais simples e intuitivo, e também tende a trabalhar muito bem para uma variedade de classes de problemas (BENGUEDDACH; NIAR; BELDJILALI, 2011), (NAKANO; ALMEIDA; STEINER, 2018). No FF as aplicações buscam o primeiro DC com capacidade disponível apto a atender seus requisitos mínimos. A busca pelo primeiro DC inicia na origem da aplicação e procura o DC alcançável com o menor número de saltos.

Uma análise do estudo de caso proposto é realizada. Para fins de avaliação, foram criadas duas topologias da seguinte forma: Topologia 1 (T1) - computação em nuvem tradicional (DCs apenas no núcleo da rede); Topologia 2 (T2) - Hierarquia de DCs (com três níveis). A avaliação traz uma relação dos diferentes tipos de aplicação com os níveis de DCs propostos, possibilitando um estudo sobre o dimensionamento da rede, podendo ser apontado onde é necessário uma maior atenção por parte dos fornecedores de serviços em nuvem.

A este respeito, os objetivos específicos abordados nesta dissertação são apresentados a seguir:

- Apresentar uma formulação matemática para otimização do provisionamento de recursos em uma hierarquia de DCs;
- Realizar uma comparação entre o modelo proposto e o FF;

- Realizar uma análise dos resultados da formulação matemática;
- Analisar os diferentes tipos de perfis de aplicação;
- Divulgação dos resultados através de publicações relevantes a área.

1.3 Organização da Dissertação

Este documento está dividido como segue:

- Capítulo 2: Neste capítulo são introduzidos os conceitos, as tecnologias e apresentação do estado da arte necessário para a compreensão do estudo realizado.
- Capítulo 3: Apresentam-se os trabalhos relacionados a proposta e aos estudos de caso desta dissertação, abordando alguns dos desafios associados a utilização da hierarquia de DC, assim como as vantagens.
- Capítulo 4: Apresenta-se a formulação matemática, explicando de forma detalhada as principais características do modelo.
- Capítulo 5: Neste capítulo são expostos os resultados obtidos através de uma análise, discutindo acerca da alocação dos diferentes tipos de aplicação.
- Capítulo 6: Este capítulo apresenta as considerações finais da proposta do trabalho, apontando as dificuldades encontradas e os possíveis desdobramentos de trabalhos futuros.

2 Revisão Literária

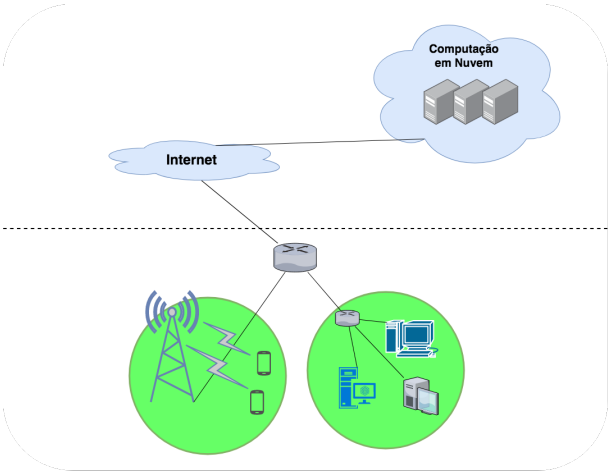
2.1 Considerações iniciais

Neste capítulo concentra-se em apresentar uma visão geral sobre os temas abordados neste trabalho com o propósito de agregar um aporte teórico. De forma sucinta, são retratados os conceitos da computação em nuvem, computação de borda, otimização linear, 5G e suas aplicações.

2.2 Computação em Nuvem

Computação em nuvem é definida em (MELL; GRANCE et al., 2009) como um modelo para permitir acesso à rede onipresente, conveniente e sob demanda a um conjunto compartilhado de recursos computacionais configuráveis (por exemplo, redes, servidores, armazenamento, aplicativos e serviços) que podem ser, rapidamente, provisionados e liberados com esforço mínimo de gerenciamento ou interação com o provedor de serviços. A figura 2 apresenta uma arquitetura da computação em nuvem.

Figura 2 – Arquitetura da Computação em Nuvem



Adaptado de (WANG et al., 2017)

Em (HAYES, 2008), os autores apontam que, para o usuário final, o controle total dos recursos computacionais tem um preço, o software deve ser instalado, configurado e atualizado a cada nova versão, infraestrutura computacional, sistemas operacionais e utilitários de baixo nível que devem ser mantidos, toda atualização no sistema operacional

desencadeia uma cascata de revisões subsequentes para outros programas, e assim por diante. A computação em nuvem além de eliminar essas preocupações também oferece vantagens em termos de mobilidade e colaboração. Em (HAYES, 2008) os autores também destacam que os incentivos para os fornecedores de *softwares* são semelhantes aqueles que motivam os usuários finais.

Esse modelo de nuvem é composto de cinco características essenciais, três modelos de serviço, e quatro de implantação, descritos a seguir. Em (RIMAL; CHOI; LUMB, 2009) os autores apontam que fundamentos como virtualização, escalabilidade, interoperabilidade, qualidade de serviço e mecanismos contra falhas são conceitos importantes quando se trata de computação em nuvem.

2.2.1 Características Essenciais

- *On-demand Self-service*: Um consumidor pode provisionar unilateralmente recursos de computação, como tempo de servidor e armazenamento em rede, conforme necessário, automaticamente, sem exigir interação humana com cada provedor de serviço.

- *Broad network access:* Os recursos estão disponíveis na rede e são acessados por meio de mecanismos que promovem o uso por plataformas heterogêneas do cliente (por exemplo, telefones celulares, *tablets*, *laptops* e *desktops*).

- *Resource pooling*: Os recursos de computação do provedor são agrupados para atender a vários consumidores usando o modelo *multi-tenent*, com diferentes recursos físicos e virtuais, atribuídos e redistribuídos dinamicamente de acordo com a demanda do consumidor. Há um senso de independência da localização em que o cliente geralmente não tem controle ou conhecimento sobre a localização exata dos recursos fornecidos, mas pode ser capaz de especificar a localização em um nível mais alto de abstração (por exemplo, país, estado ou *data center*). Exemplos de recursos incluem armazenamento, processamento, memória e largura de banda da rede.

- *Rapid Elasticity*: Os recursos podem ser provisionados e liberados de forma elástica, em alguns casos automaticamente, para escalar rapidamente mais ou menos recursos de acordo com a demanda. Para o consumidor, os recursos disponíveis para provisionamento muitas vezes parecem ilimitados e podem ser apropriados em qualquer quantidade e momento.

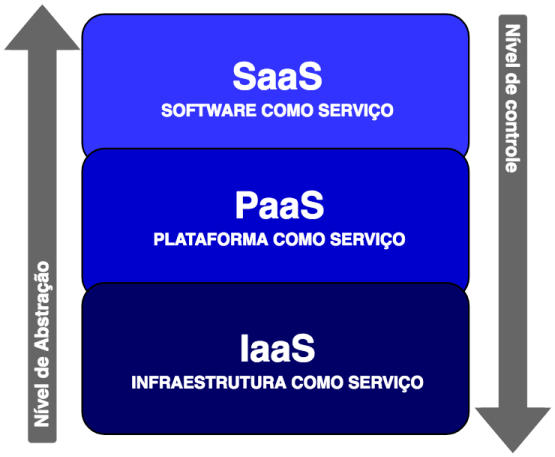
- *Measured Service*: Os sistemas em nuvem controlam e otimizam automaticamente o uso de recursos, aproveitando um recurso de medição em algum nível de abstração

apropriado ao tipo de serviço (por exemplo, armazenamento, processamento, largura de banda e contas de usuário ativo). O uso de recursos pode ser monitorado, controlado e relatado, proporcionando transparência tanto para o provedor quanto para o consumidor do serviço utilizado.

2.2.2 Modelos de Serviço

Os modelos de serviço podem ser resumidos, então, como visto na Figura 3.

Figura 3 – Modelos de Serviço em Nuvem.



Adaptado de (JANSEN; GRANCE, 2011)

- Software como serviço (SaaS - *Software as a Service*): A capacidade oferecida ao consumidor é usar os aplicativos do provedor em execução na infraestrutura de nuvem. As aplicações são acessíveis a partir de vários dispositivos do cliente por meio de uma interface, como um navegador *Web* (por exemplo, *e-mail* baseado na *Web*). O consumidor não gerencia ou controla a infraestrutura da nuvem subjacente, incluindo rede, servidores, sistemas operacionais, armazenamento ou até mesmo recursos de aplicativos individuais, com a possível exceção de configurações das aplicações específicas do usuário limitadas. SaaS pode ser referido como *Application Service Provider* (ASP). Alguns exemplos dos principais fornecedores são Oracle, Google, IBM e Microsoft.

- Plataforma como serviço (PaaS - *Plataform as a Service*): A capacidade oferecida ao consumidor é implementar na infraestrutura em nuvem os aplicativos criados ou adquiridos pelo consumidor usando determinadas linguagens de programação, bibliotecas, serviços e ferramentas suportadas pelo provedor, ou seja, é oferecido ao

desenvolvedor uma plataforma que inclua todos os sistemas e ambiente que compreendem o ciclo de vida de ponta a ponta do desenvolvimento, teste, implantação e hospedagem de aplicativos. O consumidor não gerencia ou controla a infraestrutura da nuvem subjacente, incluindo rede, servidores, sistemas operacionais, armazenamento, mas tem controle sobre os aplicativos implantados e possivelmente configurações para o ambiente de hospedagem dos aplicativos. Alguns exemplos de fornecedores de PaaS são o GAE e o Microsoft Azure.

- Infraestrutura como serviço (IaaS - *Infrastructure as a Service*): A capacidade oferecida ao consumidor é fornecer processamento, armazenamento, redes e outros recursos fundamentais de computação, nos quais o consumidor pode implementar e executar *softwares* arbitrários, que podem incluir sistemas operacionais e aplicativos. Um dos principais benefícios é o esquema de pagamento baseado no uso, o qual permite que os clientes paguem conforme as suas necessidades atuais. O consumidor não gerencia ou controla a infraestrutura de nuvem subjacente, mas tem controle sobre sistemas operacionais, armazenamento e aplicativos implantados e, possivelmente, controle limitado de componentes da rede selecionados (por exemplo, *firewalls* de *host*). Os principais exemplos são a Amazon e o GoGrid.

2.2.3 Modelos de implantação

- Nuvem privada: A infraestrutura de nuvem é provisionada para uso exclusivo por uma única organização que inclui vários consumidores (por exemplo, unidades de negócios), sem as restrições de largura de banda da rede, exposição de segurança e requisitos legais que o uso de redes públicas necessitam. Ele pode ser de propriedade, gerenciado e operado pela organização, por terceiros ou alguma combinação deles, e pode existir dentro ou fora das instalações.

- Nuvem comunitária: A infraestrutura de nuvem é provisionada para uso exclusivo por uma comunidade específica de consumidores das organizações que compartilham preocupações, tais como: missão, requisitos de segurança, políticas e considerações de conformidade. Ela pode ser de propriedade, gerenciado e operado por uma ou mais organizações da comunidade, um terceiro ou uma combinação deles, e pode existir dentro ou fora das instalações.

- Nuvem Pública: A infraestrutura de nuvem é provisionada para uso aberto pelo público em geral, na qual os recursos são dinamicamente alocados pelo provedor de serviço. Ele pode ser de propriedade, gerenciado e operado por uma organização comercial, acadêmica, governamental ou por alguma combinação deles. Existe nas instalações do provedor de nuvem.

- Nuvem Hibrida: A infraestrutura de nuvem é uma composição de duas ou mais infraestruturas de nuvem distintas (privada, comunitária ou pública) que permanecem como entidades exclusivas, mas unidas por tecnologia padronizada ou proprietária que permite portabilidade de dados e aplicativos (por exemplo, estouro de nuvem para balanceamento de carga entre nuvens).

2.3 Computação de Borda

A computação de borda foi proposta para novos tipos de serviços em nuvem, que precisam de infraestrutura computacional na borda da rede. Estudos recentes propõe a integração da computação de borda com os princípios da computação em nuvem para fornecer serviços mais complexos na borda da rede, visando reduzir a carga na nuvem centralizada e evitar gargalos e pontos únicos de falha (JARARWEH et al., 2016). A ideia central das redes de borda móvel é mover as funções de rede, conteúdos e recursos para mais perto dos usuários finais, ou seja, a borda da rede. Os recursos de rede incluem principalmente processamento, armazenamento ou armazenamento em cache e recursos de comunicação. A computação de borda é definida em (WANG et al., 2017) como: "Uma arquitetura de rede móvel que implanta e utiliza recursos de computação e armazenamento flexíveis na borda da rede móvel, incluindo a rede de acesso de rádio, roteadores de borda, *gateways* e dispositivos móveis etc". Em (WANG et al., 2017) são apresentadas algumas vantagens da computação de borda, entre elas estão: latência reduzida, redução da largura de banda e serviços de proximidade.

- Latência Reduzida: como os recursos de processamento e armazenamento estão próximos aos usuários finais, o atraso na comunicação pode ser reduzido significativamente.
- Redução da Largura de Banda: A implementação de servidores de ponta em infraestrutura de rede de borda móvel pode economizar o custo de operação em até 67% para os aplicativos que consomem muita largura de banda e para os que precisam de computação intensiva (MEHTA et al., 2016).
- Serviços de Proximidade: A arquitetura de redes de borda móvel tem grandes vantagens em fornecer serviços de proximidade, já que os servidores de borda estão mais próximos dos usuários finais e a tecnologia de comunicação D2D pode ser explorada (NUNNA et al., 2015). Portanto, a carga de tráfego na rede de acesso de rádio pode ser reduzida.

Propostas como *Fog Computing* (FOG) e *Mobile Edge Computing* (MEC) que prometem reduzir significativamente a latência ponta-a-ponta, permitindo que as novas aplicações consigam ter seus requisitos de QoS atendidos, são alguns exemplos de propostas

da computação de borda. De acordo com o Instituto Europeu de Normas de Telecomunicação (ETSI - *European Telecommunications Standards Institute*), o MEC é definido como: "A computação de borda móvel fornece um ambiente de serviços da tecnologia da informação (TI) e recursos de computação em nuvem na borda da rede móvel, dentro da rede de acesso de rádio (RAN - *Radio Access Network*) e na proximidade de assinantes móveis", (HU et al., 2015). O MEC também é equipado com técnicas de descarregamento que caracterizam a rede com baixa latência e alta taxa de dados. Em (WANG et al., 2017), o FOG é apresentado como uma plataforma projetada, principalmente, para casos de uso da IoT. Os nós FOG são massivamente distribuídos em amplas áreas, onde a principal característica é a colaboração entre vários clientes de usuários finais ou dispositivos de borda próximos do usuário para ajudar no processamento e armazenamento de dispositivos móveis.

2.4 Otimização Linear

Otimização é o ato de obter o melhor resultado sob determinada circunstância. O objetivo final de qualquer problema de otimização é minimizar o esforço necessário ou maximizar o benefício desejado. Em qualquer situação prática, o esforço ou benefício, pode ser expressado em função de certas variáveis de decisão. A otimização pode ser, então, definida como o processo de encontrar o valor máximo ou mínimo de uma função. Os métodos que buscam ótimos resultados são, também, conhecidos como técnicas de Programação Matemática (RAO, 2009).

Programação Linear (PL) ou Otimização Linear faz parte das disciplinas que compõem a Programação Matemática e é um elemento na Pesquisa Operacional (PO). A PL tem um papel duplo na programação matemática, pois os algoritmos utilizados para a sua solução podem ser de natureza combinatória (discreta) ou contínua (MACULAN; FAMPA, 2006).

A PL foi reconhecida pela primeira vez na década de 1930 pelos economistas, enquanto se desenvolviam métodos para a alocação ótima de recursos. George B. Dantzig formulou o problema geral de PL e desenvolveu o método Simplex de solução em 1947 (RAO, 2009). E, apesar de sua grande aplicação prática, o método simplex pode ter comportamento exponencial em seu número de iterações (MACULAN; FAMPA, 2006). Dantzig em 1974 forneceu os resultados teóricos e computacionais do método simplex (DANTZIG, 1951), (DANTZIG, 2016), (LENSTRA; KAN; SCHRIJVER, 1991).

Um problema de PL pode ser então definido sob a seguinte forma:

$$maximizar \quad z = \sum_{j=1}^{p} c_j x_j \tag{2.1}$$

Sujeito a:

$$\sum_{j=1}^{p} a_{i,j}x_j \leq b_i, i = 1, 2, ..., q \tag{2.2}$$

$$x_j \geq 0, j = 1, 2, ..., p \tag{2.3}$$

onde c_j, $a_{i,j}$ e b_i são dados (números reais) e x_j representa para $j = 1, 2, ..., p$, as variáveis de decisão. A função linear a ser maximizada em 2.1 é denominada Função Objetivo (FO), função econômica ou função critério. As restrições de não negatividade 2.3 são conhecidas como triviais (MACULAN; FAMPA, 2006). A seguir são apresentadas algumas definições importantes:

- Função Objetivo: é a função que relaciona o produto entre as grandezas e as variáveis que se deseja maximizar ou minimizar de acordo com o problema.
- Restrições: são os elementos que restringem o problema a seus valores possíveis de se alcançar. As equações de restrição de um problema de PL podem estar na forma de igualdade ou desigualdade.
- Solução Viável: é toda solução que satisfaz as condições do problema.
- Solução Ótima: é toda solução viável, que retorna o valor máximo ou mínimo alcançado pelo problema de maximização ou minimização, respectivamente.

As características de um problema de PL, declarado na forma padrão são (RAO, 2009):

- A FO é do tipo de minimização.
- Todas restrições são do tipo de igualdade.
- Todas as variáveis de decisão são não-negativas.

O desenvolvimento da PL tem sido classificado entre os mais importantes avanços científicos dos meados do século XX (HILLIER; LIEBERMAN, 2013). A PL, então, envolve o planejamento de atividades para obterem ótimos resultados. Desta forma, a otimização linear é usada nos mais diversos campos de pesquisa, inclusive na engenharia em áreas como balanceamento de carga, alocação de recursos, técnicas de roteamento e implantação de roteadores e sensores.

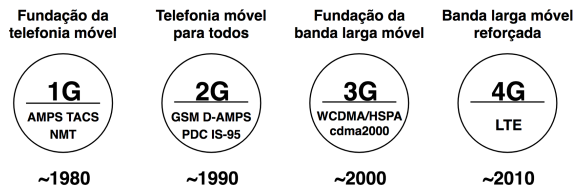
2.5 Redes Móveis de Quinta Geração (5G)

O desenvolvimento das comunicações móveis tem sido tradicionalmente visto como uma sequência de gerações nas quais, nos últimos 40 anos, o mundo testemunhou quatro destas gerações de comunicação móvel, conforme visto na Figura 4. A primeira geração (1G) de comunicação móvel, nasceu por volta de 1980, baseada em transmissão analógica com as principais tecnologias sendo AMPS (*Advenced Mobile Phone System*) e TACS (*Total Access Communication System*). Os sistemas de comunicação baseados no 1G eram limitados a serviço de voz e, pela primeira vez, pessoas comuns tinham acesso à telefonia móvel.

A segunda geração (2G) de comunicação móvel, emergiu no inicio de 1990, introduzindo a transmissão digital em *links* de rádio e com isso, a transmissão limitada de serviços de dados. Inicialmente houveram diversas tecnologias do 2G, incluindo GSM (*Global System for Mobile communication*), D-AMPS (Digital AMPS) e a PDC (*Personal Digital Cellular*). A terceira geração (3G) de comunicação móvel surgiu no inicio do ano 2000. Com o 3G, um verdadeiro passo para banda larga de alta qualidade foi dado, permitindo acesso rápido a Internet sem fio. Isso foi especialmente permitido pela evolução do 3G conhecida como HSPA (*High Speed Packet Access*).

Atualmente, estamos na quarta geração (4G) de comunicação móvel, representada pela tecnologia LTE (*Long Term Evolution*), que seguiu os passos do HSPA e promove uma melhor eficiência e experiência em termos de taxas de dados para os usuários móveis mais acessíveis (DAHLMAN; PARKVALL; SKOLD, 2018). O padrão LTE possui vários *releases*, dentre os quais o *Release* 8 foi a primeira versão a ser comercializada, com taxas de pico alcançando os 300 Mbps para *download* e 75 Mbps para *uplink* (ALI-YAHIYA, 2011).

Figura 4 – Diferentes Geração da comunicação móvel



Adaptado de (DAHLMAN; PARKVALL; SKOLD, 2018)

As discussões sobre a quinta geração de redes móveis tiveram início por volta de 2012. O 5G passou a ser usado com frequência em um contexto muito mais amplo, não apenas referindo-se a uma tecnologia de acesso de rádio especifica, mas a uma ampla variedade de novos serviços previstos para serem habilitados pelas comunicações móveis futuras (DAHLMAN; PARKVALL; SKOLD, 2018), onde o número de dispositivos pode

chegar a dezenas, ou até centenas de bilhões, quando o 5G se tornar realidade (ANDREWS et al., 2014).

A crescente proliferação de dispositivos inteligentes, a introdução de novas aplicações multimídia emergentes, juntamente com o aumento exponencial da demanda e no uso de dados sem fio (multimídia), estão criando uma carga significativa nas redes celulares existentes (AGIWAL; ROY; SAXENA, 2016). O 4G será facilmente substituído pelo 5G (GUPTA; JHA, 2015). É esperado que os sistemas sem fio 5G, com melhores taxas de dados, capacidade, latência e QoS, sejam a solução da maioria dos problemas atuais das redes celulares. O 5G promete ter características bem superiores, se comparado com seus antecessores mostrados a seguir (HOSSAIN; HASAN, 2015), (AGIWAL; ROY; SAXENA, 2016):

- Taxa de dados de 1 a 10 Gigabit por segundo (Gbps) em redes reais: isso é quase 10 vezes maior do que a taxa de dados de pico teórica da rede LTE tradicional de 150 Megabit por segundo (Mbps);
- Latência de ida e volta de 1ms (milissegundo): redução de 10 vezes do tempo de ida e volta de 10ms;
- Maior largura de banda;
- Grande número de dispositivos conectados;
- 99,99% de disponibilidade percebida: 5G prevê que a rede deve estar praticamente sempre disponível;
- Cobertura de quase 100% para conectividade "a qualquer hora e em qualquer lugar": O 5G precisa garantir cobertura completa, independente da localização do usuário;
- Redução do uso de energia em quase 90%;
- Alta duração da bateria;

O 5G não será uma melhoria incremental em relação aos seus predecessores, o objetivo é ser um salto revolucionário, onde essas capacidades visam a conectividade de alta velocidade, IoT, realidade virtual aumentada, a internet tátil, e assim por diante (SHAFI et al., 2017).

2.6 Aplicações - 5G

O advento dos sistemas celulares 5G, com disponibilidade de uma tecnologia de conectividade, que promete ser onipresente, confiável e escalável, é considerado como um

potencial motivador para o surgimento da IoT global ainda a emergir (PALATTELLA et al., 2016). O aumento da taxa de dados, a redução da latência ponta-a-ponta e a melhoria da cobertura com relação as tecnologias que antecederam o 5G, mantêm o potencial de atender, até mesmo, as aplicações mais exigentes em termos de requisitos de comunicação (PALATTELLA et al., 2016).

A IoT, também chamada de Internet de Tudo ou da Internet Industrial, é um paradigma tecnológico concebido como uma rede global de máquinas e dispositivos capazes de interagir entre si (LEE; LEE, 2015). Alguns dos principais serviços para o período de 2020+ envolvem aplicativos da agricultara à manufatura, de cidades inteligentes que exigem *smart grid*, transporte e mobilidade inteligente, medicina digital, sistemas de seguranças, entre outros (BORKAR; PANDE, 2016). A IoT visa melhorar a qualidade de vida das pessoas, economizando tempo e dinheiro (PALATTELLA et al., 2016). Algumas das áreas que a IoT vai atuar são mostradas na Figura 5.

Figura 5 – Áreas de aplicativos IoT.

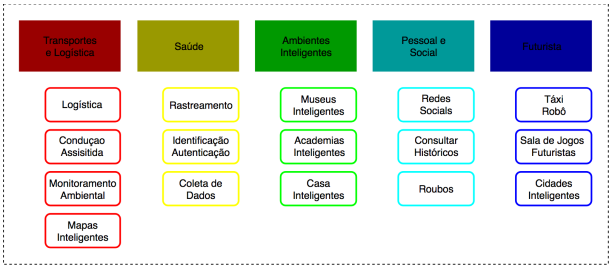
Adaptação de (DATTA; SHARMA, 2017)

A IoT pode ser apresentada como uma rede de dispositivos que se conectam à internet, como sensores, veículos, dispositivos que podem ser monitorados, detectados e controlados (HEJAZI et al., 2018) a qual suporta uma grande variedade de aplicativos com requisitos contraditórios (DATTA; SHARMA, 2017). Uma estimativa recente revela que, em média, cada indivíduo usa mais de dois dispositivos conectados à internet (AMADEO et al., 2016). As primitivas da IoT são: sensores, agregadores, canais de comunicação, utilitários externos e tomadores de decisão (VOAS, 2016).

O desenvolvimento da IoT no ambiente atual prevê muitos avanços em como vemos o mundo, e será aplicável em todas as esferas da vida (DATTA; SHARMA, 2017). Aplicações IoT são mostradas esquematicamente na Figura 6. As novas aplicações, como as que convivem em redes sociais (JAMEEL et al., 2018), vão exigir cada vez mais a confluência das tecnologias e padrões, como protocolos e comunicação, infraestrutura de rede, armazenamento e processamento, entre outras tecnologias. Alguns dos principais requisitos dessas novas aplicações são: suporte a QoS adaptável, latência significativamente

reduzida, economia de energia e alta integridade (BORKAR; PANDE, 2016).

Figura 6 – Aplicações IoT



Adaptação de (DATTA; SHARMA, 2017)

2.7 Considerações finais

Este capítulo teve como objetivo principal, apresentar as principais tecnologias e conceitos retratados neste trabalho. Tais conceitos são importantes, pois serão utilizados na proposta e no entendimento do trabalho apresentado. Desta forma, foram elencadas os conceitos, visando dar embasamento teórico para o trabalho proposto.

3 Trabalhos Correlatos

3.1 Considerações iniciais

Nos último anos, a literatura tem se concentrado na implantação de DCs para atender os novos requisitos de QoS que surgem com a próxima geração de redes móveis, o 5G. Sofisticadas abordagens de planejamento e engenharia também são necessárias serem adotadas pelos provedores de serviço para dar conta dessa heterogeneidade, inerente aos aplicativos e serviços de IoT (STURZINGER; TORNATORE; MUKHERJEE, 2017a). Dentre as soluções citadas na literatura, a mais adequada o desempenho das aplicações e serviços, é o descarregamento para os servidores em nuvem, tanto a nuvem tradicional quanto a nuvem de borda. Neste capítulo são apresentados alguns dos trabalhos relacionados ao conteúdo deste documento.

3.2 Alocação de Recursos

Em (ZHANG et al., 2018), um problema de otimização é formulado para fornecer a política ótima de estratégia de descarregamento em computação, alocação de sub-canal e potência de transmissão de *uplink* e escalonamento de recursos em computação. Os autores conseguem obter resultados satisfatórios, no qual o esquema proposto consegue, efetivamente, diminuir o consumo de energia e o tempo de resposta dos serviços, superando outros algoritmos utilizados na literatura. Em (LÓPEZ-PIRES, 2016), os autores tratam do problema de alocação de recursos em *data centers* de computação em nuvem, mapeando quais VMs devem ser hospedadas em quais máquinas físicas, conhecido como problema *Virtual Machine Placement* (VMP). É proposto um modelo de otimização de muitos objetivos, e os resultados experimentais comprovam a correção, eficácia e escalabilidade dos algoritmos propostos em diferentes ambientes experimentais.

Em (ZHARIKOV; ROLIK; TELENYK, 2018), os autores propõe um sistema de gerenciamento hierárquico multicamadas para *data centers* em nuvem, onde são combinadas técnicas tradicionais de gerenciamento de recurso com modelos hierárquicos. O modelo base é tido como dinâmico, pois pode ser utilizado para avaliar a eficácia de gerenciamento através do QoS dos serviços prestados, pelo custo dos recursos sob demanda e pelo custo de violação da Garantia do Nível de Serviço (SLA - *Service Level Agreement*). O sistema proposto permite controlar os serviços de informações de um *datacenter* em nuvem com base nos critérios de minimização de custos e desvio padrão dos parâmetros de qualidade

dos serviços fornecidos.

Em (CHEN, 2018), é proposto um método de alocação de recursos em nuvem que suporta demandas súbitas e urgentes, pois o método pode alocar, de maneira oportuna e ideal, demandas urgentes de serviços. O modelo apresentado recria as prioridades de alocação para alocar recursos com solicitações urgentes em máquinas virtuais, uma vez que o modelo de otimização multi-objetivo é estabelecido para definir a distância mínima de correspondência entre as VMs e as máquinas físicas. Os resultados obtidos pelos autores mostram que o modelo proposto consegue reduzir o número de VMs usadas enquanto melhora a utilização dos recursos para serviços urgentes.

Em (ZHANG et al., 2017), para minimizar a latência dos serviços e o consumo de energia dos dispositivos, os autores investigam a alocação de recursos multiobjetivo para sistemas MEC multiusuário adotando, como métrica, uma combinação ponderada normalizada do tempo e economia de energia alcançada pelo descarregamento de computação, em que um algoritmo baseado em ranking de complexidade é proposto. Os autores comparam a proposta com um algoritmo de alocação aleatória, denominado RanCh-HeuOff, no qual os resultado mostram que o algoritmo baseado em ranking desfruta de um desempenho quase ótimo.

Em (ALAM; ZULKERNINE; HAQUE, 2017), os autores fornecem uma abordagem para alocação de recursos para computação em nuvem, visando minimizar o custo enquanto abordam a confiabilidade como recurso mais crucial. O objetivo dos autores é maximizar a confiabilidade enquanto minimiza o custo, utilizando uma heurística como técnica de otimização. Os resultados que eles trazem apontam que a abordagem proposta fornece uma maior confiabilidade ao alocar os recursos, quando comparada com as técnicas presentes na literatura.

Em (RAHMAN et al., 2016), uma seleção de aplicações para analisar o comportamento dos serviços hospedados na nuvem e registrar as principais propriedades para avaliar o QoS, como tempo de resposta, disponibilidade, taxa de transferência, confiabilidade e latência é realizada. Os autores discutem fatores que podem impactar no serviço oferecido por diferentes empresas, dentre eles, o algoritmo implementado para gerenciamento de recursos, que é um fator relacionado ao *software* e impacta, significativamente, nos valores de QoS. Em (RUAN et al., 2018), os autores propõe um modelo de rede com descarregamento parcial, no qual parte da tarefa é processada localmente, e parte é processada no servido MEC de acordo com a meta mínima de consumo do sistema. Também é proposto uma estratégia de descarregamento de computação *uplink* distribuída e orientada para o usuário, na qual é otimizado o conjunto de decisão do descarregamento, juntamente com a competição inter-celular de oportunidades.

Em (SILVA et al., 2016), os autores investigam os benefícios de aplicar os conceitos de realocação e diferenciação de serviços ao restaurar os serviços da nuvem ótica. É

considerado um cenário de provisionamento dinâmico, em que os serviços de nuvem são divididos em classe e duas abordagens são apresentadas, uma baseada em programação linear inteira (ILP - *Integer Linear Programming*) cujo objetivo é minimizar o tempo de inatividade médio do serviço e o número de serviços em nuvem realocados, e a segunda heurística é desenvolvida para fornecer resultados de desempenhos próximos ao ILP, mas com um tempo de processamento, significativamente, menor. As propostas são avaliadas em termos das seguintes métricas: Probabilidade de bloqueio, disponibilidade, restauração e realocação média. A heurística proposta alcança resultados muito próximos do ILP, e os resultados mostraram que ambas abordagens são capazes de melhorar o desempenho médio de disponibilidade e restaurabilidade de serviços quando comparadas com técnicas convencionais baseadas em restauração.

Os trabalhos apresentados nesta seção visam otimizar a utilização dos serviços em nuvem com diversos objetivos, entre eles, podemos citar a maximização do QoS das aplicações e utilização dos recursos ou mesmo a minimização do consumo de energia. Entretanto, eles ignoram a existência da crescente demanda do tráfego proveniente das aplicações emergentes da nova geração de redes móveis, e os novos requisitos cada vez mais estritos dessas aplicações. A proposta deste trabalho díspar dos apresentados aqui, quando se preocupa em otimizar a alocação dos recursos, não apenas em um nível, mas em diversos níveis da rede, considerando os novos perfis de aplicações.

3.3 Computação de Borda

As propostas de extensão da nuvem para a borda da rede tem ganhado bastante espaço e força no meio acadêmico e industrial. Em (ARAL; BRANDIC, 2017), é proposto um modelo de rede Bayesiana de parâmetros relacionados a QoS para estimar o nível de disponibilidade de uma máquina virtual (VM - *Virtual Machine*) na infraestrutura de borda. O modelo é comparado com outros métodos de aprendizado de máquina no qual os resultados experimentais mostram que o método proposto pode identificar as VMs que satisfazem os objetivos de disponibilidade definidos pelo usuário com até 94% de precisão.

Em (FORD et al., 2017), os autores buscam otimizar o provisionamento de recursos para criar um MEC distribuído, que irá oferecer oportunidades para latências mais baixas e maior resiliência enquanto atende a demanda máxima dos usuários finais, que é responsável pela interrupção do serviço que pode ocorrer devido à mobilidade do usuário entre as áreas de serviço de diferentes *data centers*, buscando, periodicamente, minimizar os custos de rede. Os autores conseguem demonstrar uma redução de 75% na redundância de DCs da topologia atual dos operadores, mantendo o mesmo nível de resiliência, sendo possível devido a distribuição da carga por muitos DCs de nuvem móvel.

Em (JIJIN et al., 2017), os autores concentram seus esforços para resolver o pro-

blema de atribuição de tarefas em pontos de acesso FOG virtuais (v-FAPs - *Virtual FOG Access Point*). É formulado um problema de otimização para encontrar a atribuição ideal de tarefas com o objetivo de equilibrar as cargas nos nós de serviço, minimizando os custos máximos de recursos usados. Os resultados mostram que o modelo proposto pode alcançar cargas significativas mais equilibradas e menores taxas de falha, em comparação com um esquema de custo mínimo guloso.

Os autores em (VELASCO; RUIZ, 2018), analisam os componentes básicos de uma infraestrutura de Fog altamente distribuída e densa para fornecer serviços unificados e econômicos para redes 5G, como NVF, MEC e serviços para terceiros, como cidades inteligentes e IoT. Os autores apontam como principais benefícios a implantação dinâmica de novos serviços distribuídos de baixa latência. Em (GONZALEZ et al., 2016), os autores exploram o estado da arte da FOG e seus predecessores, analisando como as ideias e conceitos evoluíram, quais as relações entre eles e qual o papel da computação em nuvem na computação de borda.

Em (NGUYEN; HUH, 2018), é apresentada uma nova estrutura de simulação, chamada ECSim++, baseada no simulador OMNeT++ e no *framework* INET que permite a exploração de problemas relacionados a computação, armazenamento em cache, comunicação ou eficiência energética na computação de nuvem de borda. O ECSim++ é um simulador de eventos discretos desenvolvido na linguagem de programação C++ que suporta, totalmente, ambientes de computação de nuvem de borda. Os autores apontam como principais vantagens do ECSim++ o gerenciamento de serviços em nuvem na borda da rede, um novo protocolo para gerenciar serviços, caches no nó de borda e um módulo de energia.

Nesta seção, os trabalhos citados apresentam a computação de borda e mostram que ela se torna crucial para atender as novas demandas de aplicações. Entre os benefícios trazidos pela computação de borda, podemos citar as minímas latências e a atenuação do tráfego dos *links* de *backhaul*, evitando gargalos e pontos únicos de falha. Com a sua proposta, a computação de borda se torna um tema atual, entretanto, os trabalhos apresentados nesta seção não consideram a coexistência da computação de borda e de nuvem no mesmo cenário, fator que pode ser crucial para o melhor aproveitamento dos recursos disponíveis na rede. Desta forma, este trabalho difere dos demais apresentados, pois trata da coexistência da computação de borda e da computação de nuvem a qual visa otimizar a utilização dos recursos disponíveis na rede de modo a anteder os diferentes tipos de serviços.

3.4 Hierarquia de Data Centers

Um modelo para o provisionamento de recursos em redes de área metropolitana em uma arquitetura híbrida de nuvem e FOG, procurando minimizar o custo operacional total de provisionamento das demandas de tráfego IoT, bem como fornecer uma estrutura para realocação dinâmica de *laightpah* na rede de área metropolitana (MAN - *Metropolitan Area Network*) é proposto em (STURZINGER; TORNATORE; MUKHERJEE, 2017a). O modelo apresentado encontra o provisionamento de recursos ideal, apontando onde as aplicações serão alocadas, se na nuvem ou na FOG, assim demonstrando quais perfis de aplicação e parâmetros topológicos têm os efeitos mais significativos sobre os componentes de custos individuais. Os autores, em (STURZINGER; TORNATORE; MUKHERJEE, 2017a), apontam como resultados que, à medida que a latência aumenta, aplicativos de maior complexidade podem, ser transferidos para a nuvem a fim de reduzir custos.

Em (DALLA-COSTA et al., 2017), é proposto um orquestrador de virtualização de funções de rede (NFV - *Network Functions Virtualization*) para ambientes sem fio. O orquestrador é capaz de decidir entre várias composições possíveis de NFV, quais são as mais adequadas para cada situação. Os autores mostram que, utilizando o orquestrador proposto, há uma diminuição significativa da largura de banda consumida na rede de *fronthaul* na ordem de 70%, pois as funções de banda básica da RAN centralizada (C-RAN - *Centralized*-RAN) dinâmica são mantidas perto das estações bases (BS - *Base Station*), enquanto funções de alto nível são implementadas em nuvens distribuídas hierarquicamente.

Em (HONG, 2017), é proposto um ecossistema de computação em nuvem e FOG. Um *testbed* para os algoritmos propostos para melhor atender a demanda por serviços com base na latência é realizado para três cenários: (i) disseminação de conteúdo em redes desafiadas, (ii) Fog *Crowd-sourced* e (iii) IoT analítica programável. Os autores apontam que os algoritmos propostos superam os tradicionais em 30%, 20% e 89%, para os três cenários de uso, respectivamente. Em (ALNAZIR et al., 2017), os autores utilizam o *software* Cloud-Sim para realizar um estudo do desempenho da computação em nuvem com DCs distribuídos. Os autores conseguem constatar que existe uma redução do custo total e no tempo de resposta, quando são utilizados DCs distribuídos.

Em (KHAN; FREITAG, 2017), é apresentada um protótipo para nuvem de borda colaborativa extensível, na qual os recursos de dispositivos domésticos de baixa capacidade serão compartilhados entre os usuário, e desta forma, a elasticidade dos recursos e prestação de serviços na borda da rede pode ser obtida. No entanto, a nuvem de borda colaborativa também aproveitará os serviços da computação em nuvem tradicional para melhorar o desempenho dos serviços oferecidos. No modelo proposto em (KHAN; FREITAG, 2017), os principais recursos da nuvem são fornecidos pelos dispositivos distribuídos na borda. Em (JARARWEH et al., 2016), é proposto um modelo hierárquico composto

de servidores MEC e uma infraestrutura de *Cloudlets*. O objetivo do modelo é aumentar a área de cobertura para os usuários móveis, em que os usuários podem realizar seus serviços solicitados com custos mínimos, em termos de potência e atraso. Os resultados mostram que o modelo proposto é superior ao da literatura com que foi comparado.

Nesta seção os trabalhos apresentados trazem a computação de borda como uma extensão da nuvem. Os autores não se preocupam mais em tratar da computação de borda ou em nuvem individualmente, entretanto, vale ressaltar que o foco ainda é a computação de borda, sendo a nuvem tradicional deixada em segundo plano. Diferente dos trabalhos apresentados, este produção busca otimizar a utilização dos recursos da computação de borda e nuvem, igualmente, sem fazer distinção de quem oferece o serviço, desde que este atenda os níveis de QoS exigidos pelas novas aplicações que estão surgindo.

3.5 Considerações finais

Neste capítulo, são apresentados os trabalhos correlatos a este. É visto que os temas apresentados neste trabalho, como alocação de recursos e computação em nuvem, ainda são temas bastantes discutidos na literatura, em suas mais diversas formas. Assim, os trabalhos apresentados visam atender os requisitos cada vez mais restritos das novas aplicações, através de várias técnicas de otimização para uma melhor alocação de recursos na computação em nuvem.

A proposta deste trabalho, diferente das propostas apresentadas anteriormente, considera a coexistência de DC em diversos níveis da rede sem focar, apenas, na borda ou na computação em nuvem tradicional, como em (KHAN; FREITAG, 2017). A Hierarquia de DCs é proposta visando solucionar problemas, como congestionamento no núcleo da rede e obtenção de latências menores que 10ms, ocasionados pelas novas demandas de serviços. O modelo proposto, então, visa dimensionar e otimizar o provisionamento de recursos nos estudos de caso, apresentados nos capítulos posteriores.

4 Formulação Matemática

4.1 Considerações iniciais

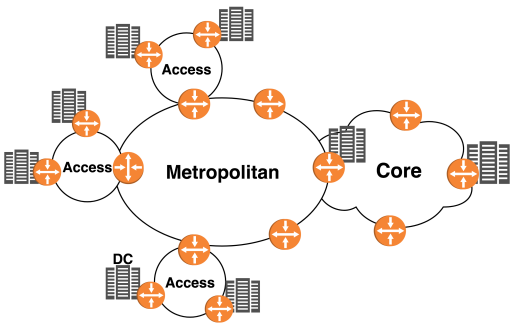
Neste capítulo, é descrita a formulação matemática utilizada. O modelo proposto otimiza o uso dos recursos de rede em um cenário, hierarquicamente, distribuído com três níveis, aumentando o uso dos recursos computacionais utilizados e o número de aplicações que são atendidas. Neste capítulo são apresentados os parâmetros e variáveis, a função objetivo e as restrições do modelo.

4.2 Topologia e Variáveis

O modelo consiste em dois componentes principais: um conjunto de aplicações R e um conjunto de nós C, representando a topologia de rede. O conjunto de aplicações R será a entrada do modelo, e a saída será constituída por duas variáveis: $X^r_{i,j}$, que representa o percursos feito por uma aplicação do conjunto R e $P_{r,i}$, que representa o nó escolhido por cada aplicação para processamento e/ou armazenamento.

A topologia proposta, composta de três níveis de DC, está representada na Figura 7. O conjunto de nós C que compõe a topologia é divido em três tipos, de acordo com seu posicionamento na rede, assim tem-se: nós de borda ou *datacenters* de borda (DCB), nós de metro ou *datacenters* regionais (DCR) e nós de núcleo ou *datacenters* de núcleo (DCN). A capacidade computacional é um fator que diminui com o distanciamento dos DCs do núcleo da rede, sendo assim, os DCs de borda com menor capacidade computacional.

Figura 7 – Rede Hierarquicamente Distribuída em Três Níveis.



O conjunto C é dividido em outro sub-grupo composto de dois tipos: nós de roteamento e nós de processamento. Os nós de roteamento servem, apenas, para encaminhar as aplicações pela topologia, até encontrarem um DC capaz de processá-las, e os nós de processamento, por sua vez, são os DCs, destinos finais das aplicações. Entretanto, um nó de processamento também pode servir como um nó de roteamento.

O conjunto C é composto por seis parâmetros, definidos a seguir.

- $DC_n = 1$ se o nó n for um DC.
- CP_n é a quantidade de unidades de processamento (UP) em que o nó n possui.
- ST_n é a quantidade de unidades de armazenamento (US) em que o nó n possui.
- $T_{i,j} = 1$ se o nó i está ligado ao nó j.
- $Dist_{i,j}$ é a distância do link que liga os nós i e j.
- $Wl_{i,j}$ é o total de comprimento de ondas existente na fibra que liga os nós i e j.
- Vf é a velocidade dos dados na fibra.

O conjunto de aplicações R é composto por sete elementos, definidos a seguir:

- ID_r representa o *id* da aplicação r;
- Tp_r representa o tipo da aplicação r;
- SRC_r representa a origem da aplicação r;
- P_r representa quantas UP são necessárias para processar a aplicação r;
- S_r representa quantas US são necessárias para processar a aplicação r;
- W_r representa quantas unidades de onda a aplicação r vai ocupar no link;
- D_r representa o atraso máximo permitido da aplicação r;

O conjunto de aplicações R, conforme apontadas em (STURZINGER; TORNATORE; MUKHERJEE, 2017b), estão representadas na Tabela 1, juntamente com seus parâmetros de processamento, armazenamento e latência. Nesta formulação considera-se, apenas, o tempo de ida da aplicação para o DC como atraso.

O modelo proposto utiliza duas variáveis do tipo binário para representação do fluxo e alocação das aplicações.

Tabela 1 – Perfis de Aplicação

Aplicação	Atraso (ms)	UP (CPU/Mbps)	US (Megabit)
Realidade Virtual (RV)	10	0.03	100
Automação Industrial (AI)	20	0.009	1
Backup de Dados (BD)	1000	0	1
Smart Grid (SG)	50	0.007	0.4
Casa Inteligente (CI)	60	0	0.001
Médica (M)	40	0.020	2
Monitoramento Ambiental (MA)	1000	0.002	1
Internet Tátil (IT)	1	0.005	200

Fonte: (STURZINGER; TORNATORE; MUKHERJEE, 2017b)

- $X_{i,j}^{r} = 1$ se a aplicação r saiu do nó i para o nó j, representando o fluxo da aplicação até o seu destino.
- $Ar, i = 1$ se a aplicação r foi alocada no nó i.

4.3 Função Objetivo

O problema foi modelado através de uma função multi-objetivo, composta por três objetivos, que visam maximizar o total de aplicações alocadas e minimizar o total de recursos utilizados. A função, então, é descrita em 4.1:

$$Minimize \quad Z = Aloc_A + Use_P + Path \tag{4.1}$$

$Aloc_A$, $Aloc_P$ e $Path$, representam o total de aplicações alocadas, o total de recursos computacionais ocupados (processamento e armazenamento) e o total de *links* de fibra usados, respectivamente. Os objetivos são descritos, individualmente, abaixo:

$$Aloc_A = T_A - \sum_{r \in R} \sum_{n \in C} A_{r,n} \tag{4.2}$$

A equação 4.2 representa o total de aplicação que não foram alocadas. Onde TA é o total de aplicações que existe naquele instante de tempo, e o somatório, representa o total de aplicações que foram alocadas. Assim, minimizando 4.2, maximiza-se o número de aplicações alocadas.

$$Use_P = T_P - \left[\left(\sum_{r \in R} \sum_{n \in C} A_{r,n} * P_r \right) + \left(\sum_{r \in R} \sum_{n \in C} A_{r,n} * S_r \right) \right] \tag{4.3}$$

A equação 4.3, representa o total de recursos computacionais (processamento e armazenamento) que não foram utilizadas. Onde TP é o total de recursos na rede, e os

somatórios, representam o total de UP e US que foram utilizadas, respectivamente. Sendo assim, minimizando a 4.3, maximiza-se a utilização de recursos computacionais.

$$Path = \sum_{r \in R} \sum_{i \in C} \sum_{j \in C} X_{i,j}^{r} \tag{4.4}$$

A equação 4.4, representa o total dos recursos *links* de fibra que foram utilizados, em que o somatório representa o total de saltos que uma aplicação fez até chegar no seu destino final. Assim, minimizando o número de saltos da aplicação, é minimizado o uso dos recursos da fibra.

4.4 Restrições

A função objetivo garante que o máximo de aplicações sejam alocadas em um melhor uso dos recursos de rede. Entretanto, se faz necessário um conjunto de regras para garantir que as restrições da rede e das aplicações sejam respeitadas, por exemplo, a latência minima aceitável de uma aplicação, ou mesmo a quantidade máxima de UP que um DC pode suportar, não seja ultrapassada. Sendo assim, as equações 4.5, 4.6, 4.7 e 4.8 garantem que os parâmetros da rede sejam respeitados.

$$\sum_{r \in R} (A_{r,n} * P_r) \leq CP_n, \forall n \in C \tag{4.5}$$

$$\sum_{r \in R} (A_{r,n} * S_r) \leq ST_n, \forall n \in C \tag{4.6}$$

$$\sum_{i \in C} \sum_{j \in C} \frac{X_{i,j}^{r} * Dist_{i,j}}{Vf} \leq D_r, \forall r \in R \tag{4.7}$$

$$\sum_{r \in R} X_{i,j}^{r} * W_r \leq Wl_{i,j}, \forall (i \in C, j \in C) \tag{4.8}$$

As equações 4.5 e 4.6, garantem que um DC não terá mais que sua capacidade preenchida, ou seja, no máximo CP_n UP e ST_n US vão ser alocadas naquele nó, respectivamente. Nós, somente de roteamento, possuem $CP_n = 0$ e $ST_n = 0$, sendo assim, não é possível alocar aplicações neles. 4.7 garante que uma aplicação terá sua latência minima respeitada, sendo assim, nenhuma aplicação alocada terá seu QoS influenciado por atrasos maiores que o requerido. 4.8 restringe que uma fibra seja utilizada infinitamente, sendo sua capacidade limitada.

As equações de 4.9 à 4.16, garantem o percurso da aplicação, fazendo com que ela percorra caminhos válidos e evite congestionamento, limitando o número de aplicações que transitam na rede.

$$\sum_{kinC} X^r_{n,k} + P_{r,n} = 1, \forall(r \in R, n \in C, SRC_r = n, DC_n = 1, T_{n,k} = 1) \tag{4.9}$$

$$\sum_{k \in C} X^r_{n,k} = 1, \forall(r \in R, n \in C, SRC_r = n, DC_n = 0, T_{n,k} = 1) \tag{4.10}$$

$$\sum_{k \in C} X^r_{n,k} \leq 1, \forall(r \in R, n \in C, SRC_r \neq n, DC_n = 1) \tag{4.11}$$

$$\sum_{k \in C} X^r_{n,k} - \sum_{i \in C} X^r_{i,n} = 0, \forall(r \in R, n \in C, SRC_r \neq n, DC_n = 0) \tag{4.12}$$

$$\sum_{k \in C} X^r_{j,k} = 0, \forall(T_{j,k} = 0, r \in R) \tag{4.13}$$

$$X^r_{j,k} + X^r_{k,j} \leq 1, \forall(r \in R, j \in C, k \in C, j \neq k) \tag{4.14}$$

$$\sum_{n \in C} A_{r,n} \leq 1, \forall r \in R \tag{4.15}$$

$$A_{r,j} - \sum_{k \in C} X^r_{k,j} \leq 0, \forall(r \in R, j \in C, DC_j = 1) \tag{4.16}$$

As equações 4.9 e 4.10, tratam da aplicação no seu nó de origem. 4.9, considera que o nó de origem da aplicação r é um DC, assim, a aplicação pode ser alocada no nó de origem ou seguir para um nó vizinho. 4.10 considera que o nó de origem da aplicação r não é um DC, nesse caso, a aplicação, necessariamente, precisa seguir para um nó vizinho.

As equações 4.11 e 4.12 tratam da aplicação em um nó de transição, sendo assim, a aplicação r não está mais na sua origem. 4.11 considera que o nó de transição n é um DC, então a aplicação pode ser alocada nele ou seguir para um nó vizinho. 4.12 considera que o nó de transição n não é um DC, logo a aplicação precisa seguir para um nó vizinho.

A equação 4.13 garante que uma aplicação r só vai de um nó i para um nó j se existir um *link* entre os nós, impedindo que um *link* se mova para um nó que não esteja conectado ao seu nó atual. 4.14 garante que a aplicação r não vai ficar em *loop* entre dois nós. Assim, uma aplicação que foi do nó i para o nó j não poderá usar aquele *link*, novamente.

A equação 4.15 garante que uma aplicação r pode ser alocada no máximo em um DC, e a 4.16 garante que o DC que a aplicação está alocada precisa ser o final do percurso. Sendo assim, uma aplicação não pode ser alocada em um DC que não seja o último nó visitado.

4.5 Considerações finais

Neste capítulo foi apresentada a formulação matemática do modelo proposto para uma rede, hierarquicamente, distribuída de três níveis. Foram apresentados os parâmetros e as variáveis utilizadas, a função objetivo e as restrições do modelo. O modelo proposto consiste na maximização do uso dos recursos de rede assim, como na maximização das aplicações alocadas.

5 Estudos de Caso e Resultados

5.1 Considerações iniciais

Neste capítulo serão apresentados os cenários desenvolvidos, os parâmetros utilizados (aplicações, capacidades dos data centers, etc) bem como os resultados obtidos. Para fins de comparação, foi desenvolvido um *benchmark* baseado no *First-Fit* (onde a primeira aplicação a chegar é a primeira a ser atendida). Os resultados são fragmentados em dois grupos: Utilização dos recursos e Alocação das aplicações. Estes dois grupos, apesar de serem apresentados separadamente fazem parte de um mesmo conjunto, sendo necessário uma análise de todas as partes para uma conclusão sólida.

5.2 Estudo de Caso

No presente trabalho, são utilizadas duas topologias de DCs. A primeira representa a computação em nuvem tradicional, com DCs apenas no núcleo da rede, denominada T1, oferecendo serviços de processamento e armazenamento. A segunda topologia representa uma hierarquia de DCs com três níveis, denominada T2, em que há a coexistência da computação em nuvem tradicional com DCB e DCR, servindo como uma extensão da nuvem tradicional, oferecendo serviços de processamento e armazenamento mais próximo dos usuários finais.

Os cenários apresentados neste trabalho visam possibilitar um estudo aprofundado do modelo proposto em diferentes topologias de DCs. A tabela 2 e 3 apresentam os parâmetros utilizados em T1 e T2:

Tabela 2 – Parâmetros das Topologias

	T1	T2
Total de Nós	14	14
Total de DCs	1	6
Total de UPs	2600	2600
Total de USs	X	X
Total de Links	19	19

Além do modelo proposto, mencionado anteriormente, foi implementado um *benchamrk* para fins de comparação. Este algoritmo, baseado no FF, visa alocar as aplicações nos primeiros DCs disponíveis, não havendo preocupação se os mesmos são adequados

Tabela 3 – Parâmetros dos DC

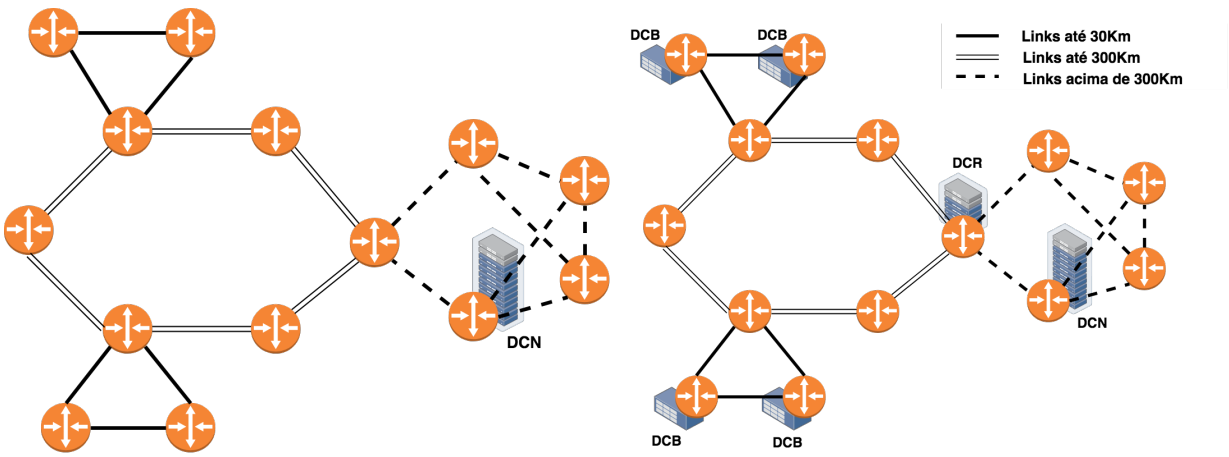
	T1			T2		
	Total	UPs	USs	Total	UPs	USs
DCN	1	1500	2000	1	2600	4000
DCR	1	500	1000	0	0	0
DCB	4	600	1000	0	0	0

e/ou, se esta alocação maximiza a utilização dos recursos em todos os DCs. Ambas abordagens utilizam o algoritmo de menor caminho (menor número de saltos) no processo de procura dos DCs.

Um conjunto de aplicações R, apresentadas em 1, é gerado aleatoriamente utilizando uma distribuição linear através de um *script* desenvolvido na linguagem de programação MatLab. Para validação estatística são gerados trinta conjuntos de aplicações. Cada conjunto sorteia as aplicações até que o total de UPs seja igual a 2600, quantitativo suportado por ambas as topologias. Assim, o conjunto R, serve como entrada para os testes realizados pelo modelo proposto e pelo FF. O *script* gerador das aplicações e o FF foram executados com auxílio do *software* MatLab ®(*MatrixLaboratory*). O modelo apresentado foi executado com auxílio do *software IBM CPLEX Optimization Studio* ®.

Abaixo, apresenta-se os cenários descritos em 2. As duas abordagens foram submetidas trinta vezes aos testes variando os conjuntos de aplicações dados como entrada para melhor validação estatísticas. O intervalo de confiança utilizado é de 95%.

Figura 8 – Topologias de DCs



(a) Topologia 1 - Computação em Nuvem (b) Topologia 2 - Hierarquia de DC

Fonte: o autor.

5.3 Resultados

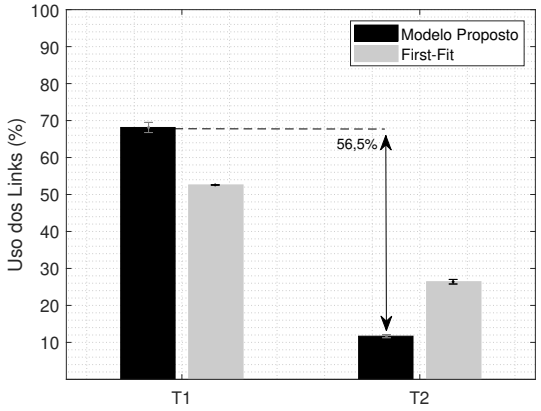
5.3.1 Utilização de Recursos

O processo de alocação otimizada busca distribuir as aplicações nos DCs e, para isso, considera em sua formulação os *links* que devem ser utilizados. Tal fato garante que a utilização dos *links*, no cenário proposto, atrelado a uma topologia hierarquicamente distribuída, seja superior as outras propostas.

A utilização dos links na T2 utilizando o modelo proposto é de 11.68%, sendo 14.72% superior que o segundo com menor utilização (T2 com FF). Esse baixo número de deslocamentos das aplicações até o núcleo da rede, ocasionado pela melhor distribuição dos recursos computacionais em T2 com modelo proposto, gera melhores fluxos pelo *backhaul*, visto que o congestionamento que seria causado pelo excesso de aplicações que precisariam trafegar até os DCN é evitado.

Importante destacar que a T1 com FF obtém uma menor utilização dos *links* de *backhaul*, que seu equivalente (T1 com modelo proposto). Esta maior concentração não reflete em uma melhor utilização dos *links*, visto que uma sobrecarga de aplicações tornaria o modelo não escalável. Desta forma, verifica-se que T1 com FF utilizou 52.59% dos *links* e T1 com modelo 68.13%, conforme é mostrado na Figura 9.

Figura 9 – Utilização dos Links (%)



Fonte: o autor.

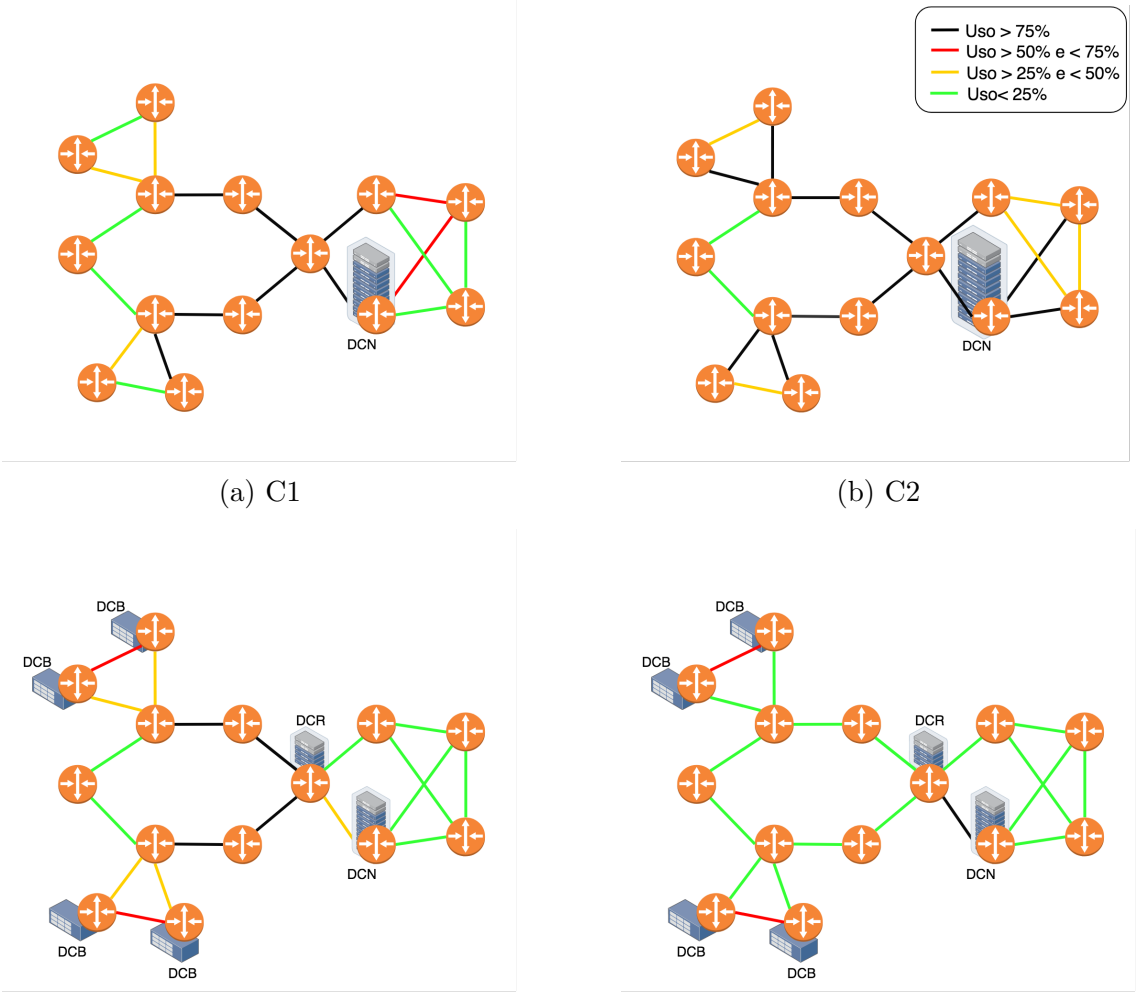
Adicionalmente pode-se observar que T2 obteve um menor uso dos *links*, ocasionando, em média, uma redução de 41.3%. Tal fato é atribuído aos recursos computacionais, visto que os mesmos são distribuídos de maneira hierárquica, acarretando um menor deslocamento para o núcleo da rede, enquanto que, em T1 as aplicações necessitam percorrer os *links* de *backhaul* até o núcleo em busca de recursos disponíveis.

Individualmente, 63% (modelo proposto) e 36% (FF) dos *links* em T1 foram ocupa-

dos em mais de 75% de sua capacidade. Em T2 ocorreu o inverso, 84% (modelo proposto) e 42% (FF) obtiveram menos de 25% da sua capacidade utilizada, como pode ser observado na Figura 10. A menor utilização dos *links* de *backhaul* é realizada quando se aproxima os recursos computacionais para a borda da rede, minimizando a chance de congestionamentos.

A baixa utilização no uso dos *links* possibilita o aumento do número de aplicações que irão circular pela rede e, com isso, é gerada uma diminuição na necessidade da implantação de novos *links* para atender o fluxo excedente. Esta melhor distribuição ocasiona uma maior utilização dos *links* presentes na borda da rede, fato desejável, para um maior aproveitamento atendimento das aplicações.

Figura 10 – Uso Individual dos Links (%)



(a) C1 (b) C2

(c) C3 (d) C4

Fonte: o autor.

É importante salientar que, o frequente uso dos *links* não significa, necessariamente, um maior uso dos recursos computacionais disponíveis. Se faz necessário uma estratégia de

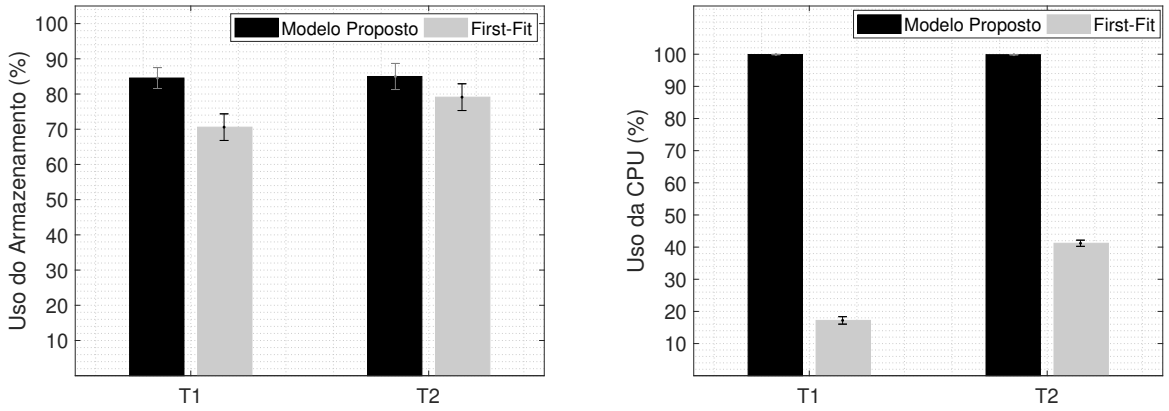
alocação de aplicações melhor, especialmente em cenários hierarquicamente distribuídos, no qual a alocação, de forma desordenada, pode gerar resultados insatisfatórios.

A modelagem proposta possibilita identificar onde é preciso uma atenção maior dos operadores de rede para alocação de mais ou menos recursos computacionais, a fim de atender a nova e crescente demanda de aplicações com estritos níveis de QoS. A quantidade de dados que está sendo gerada, e precisa ser apenas armazenada, também tem um impacto significativo no uso desses recursos, podendo influenciar diretamente no planejamento e dimensionamento destes.

Na Figura 11, são apresentados a utilização dos recursos de armazenamento e processamento. T1 com modelo proposto consegue utilizar 99.9% e 87.54% dos recursos de processamento e armazenamento, respectivamente. Já em T2 com modelo proposto, os percentuais são de 99.9% e 85%. Para o FF, em T1 obteve-se 17.2% e 70.5% do uso dos recursos de processamento e armazenamento, respectivamente. Em T2, o uso é de 41.2% e 79.1%.

Observa-se que o modelo proposto consegue maximizar o uso dos recursos de armazenamento e processamento. Destaca-se que com uma distribuição adequada, esses recursos disponíveis alcançam baixas taxas de ociosidade, diferente do FF que, devido sua distribuição de forma gulosa das aplicações, os recursos de processamento se encontram sub-utilizados, em média 70% desses recursos permanecem ociosos. Com uma melhor estratégia é possível realizar um provisionamento adequado e um excelente dimensionamento da alocação destes recursos, obtendo uma disponibilidade ideal para a infinidade de aplicações que estão surgindo.

Figura 11 – Utilização dos Recursos Computacionais (%)



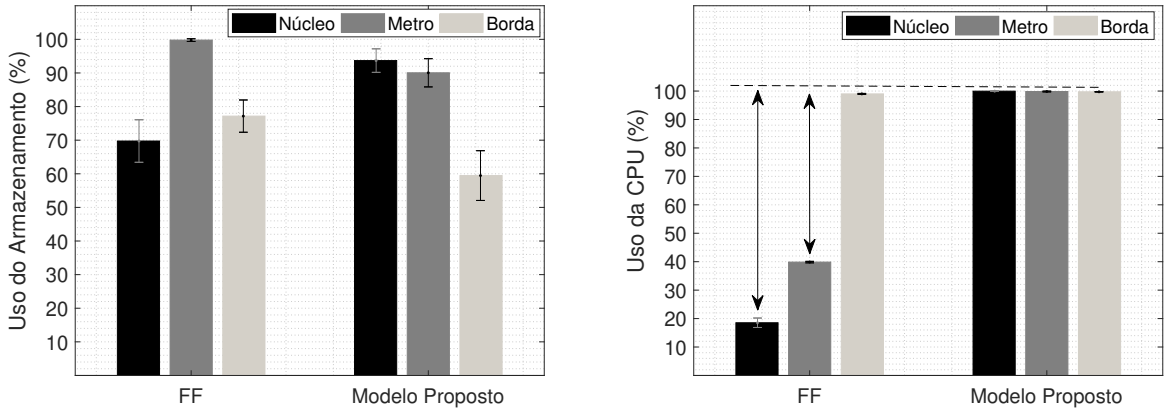
(a) Recursos de Armazenamento (%) (b) Recursos de Processamento (%)

Fonte: o autor.

O comportamento do uso dos recursos computacionais em T2 são detalhados na

Figura 12. Observa-se que o uso dos recursos de processamento no modelo proposto nos três níveis alcançaram valores próximos de 100%, enquanto que em FF, apenas os recursos disponíveis na borda conseguiram taxas similares. Isso ocorre pelo fato de muitas aplicações que necessitam de latências mínimas, não conseguirem alcançar o DCN ou mesmo o DCR.

Figura 12 – Utilização dos Recursos Computacionais na Hierarquia de DCs (%)



(a) Recursos de Armazenamento (%) (b) Recursos de Processamento (%)

Fonte: o autor.

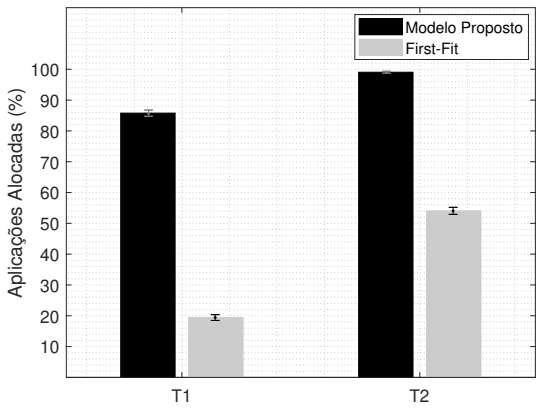
Nota-se então, que a utilização da hierarquia de DCs atrelada a uma adequada estratégia de alocação de aplicações/recursos computacionais, quando comparada as outras estratégias apresentadas, é bem superior, sendo escalável, e conseguindo distribuir melhor os recursos computacionais.

5.3.2 Alocação das Aplicações

No que concerne ao número de aplicações alocadas, o modelo proposto aderente à hierarquia de DCs também ganha dos demais cenários. T2 com o modelo proposto, alocou em média 99% das aplicações do conjunto R. Esse alto desempenho é proveniente dos recursos computacionais estarem distribuídos na rede de forma hierarquia, conseguindo atender a diversificada gama de aplicações com os mais variados tipos de restrições de QoS atrelada a uma boa estratégia de alocação das aplicações. A figura 13 apresenta o total de aplicações alocadas em cada um dos cenários.

T1 com o modelo proposto, consegue alocar 85.7% das aplicações disponíveis na rede. Já em T2 com FF, o total de aplicações alocadas é de 31.7% menor, sendo este de 54%. Observa-se então, que a hierarquia de DCs sem uma estratégia otimizada de alocação não consegue atender o conjunto de aplicações em sua totalidade. Sendo assim, se faz necessário atrelar a hierarquia de DCs com uma eficaz técnica de alocação.

Figura 13 – Medias das aplicações alocadas (%)



Fonte: o autor.

O baixo número de aplicações alocadas em T2 com FF, ocorre devido as várias aplicações com níveis estritos de latência não conseguirem ser alocadas na borda da rede, tendo em vista que outras aplicações com níveis menos restritos, já preencheram os recursos disponíveis.

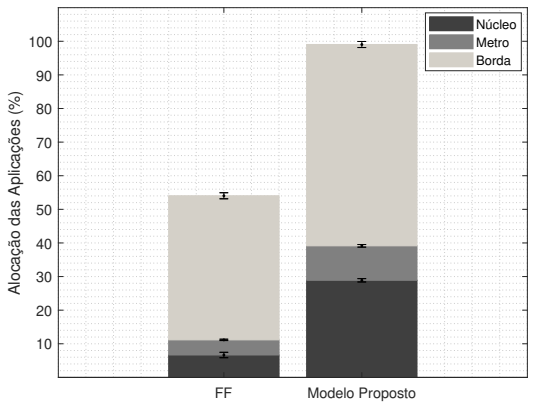
Essa alocação de aplicações em T2 pode ser melhor analisada quando observada a Figura 14. Nela são apresentadas as aplicações alocadas nos diferentes níveis de DCs disposto na rede. Percebe-se que os DCB são utilizados em suas taxas máximas, em ambas abordagens (modelo proposto e FF).

Entretanto, os DCR e DCN dependem das aplicações estarem aptas a se deslocarem até eles para serem processadas e/ou armazenadas. Através dos resultados é possível perceber a necessidade de classificar e priorizar os diferentes perfis de aplicações. Essa necessidade se torna ainda mais clara quando percebe-se que a taxa de utilização dos recursos de processamento e armazenamento dispostos no núcleo da rede chegam a elevada diferença de 81% e 24%, respectivamente. O mesmo vale para os recursos de processamento e armazenamento disponíveis na rede metro com diferenças de 60% e 13%, respectivamente.

Observa-se então, que sem uma caracterização do comportamento dos diferentes perfis de aplicações que irão ser alocadas na borda da rede, o potencial da hierarquia de DCs não é usado em seu máximo. Caracterizando essas aplicações, é possível tomar decisões de quais serão priorizadas para receber atendimento nos DCB. A complexidade computacional e os estritos tempos de latência são utilizados para caracterizar os diferentes perfis de aplicação neste trabalho.

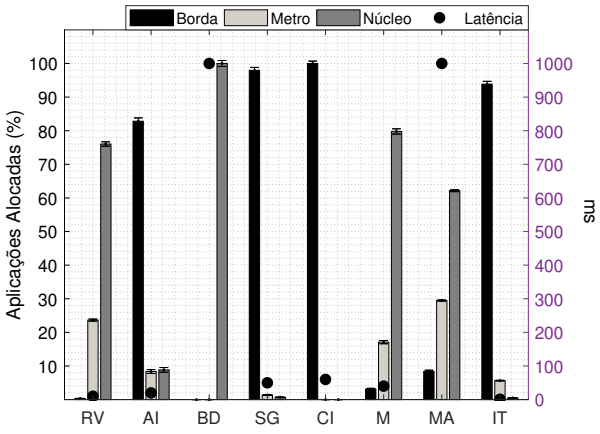
A Figura 15 apresenta o comportamento das aplicações em T2 com o modelo proposto, no qual nota-se que, conforme o tempo de latência e complexidade computacional aumentam, as aplicações migram em direção ao núcleo.

Figura 14 – Aplicações Alocadas por níveis de DCs



Fonte: o autor.

Figura 15 – Caracterização das aplicações



Fonte: o autor.

Os resultados obtidos mostram que aplicações como *Backup de Dados* que possuem maiores tempos de latência se concentram no núcleo da rede, enquanto aplicações como Internet Tátil, com os menores tempos, se localizam na borda da rede. Com essa caracterização do comportamento das aplicações alocadas, é possível que os operadores de redes criem estratégias para maximizar o uso dos recursos disponíveis de forma eficiente, atendendo ao maior número de aplicações.

5.4 Considerações finais

Neste capítulo foram apresentados e discutidos os resultados obtidos através deste trabalho. O desempenho superior do modelo proposto em relação ao FF, assim como o da hierarquia de DCs, em relação a computação em nuvem tradicional, são discutidos.

Destaca-se a importância de uma eficiente estratégia para alocação dos diferentes perfis de aplicação, assim como o melhor uso dos recursos disponíveis. Tais fatos apresentados possibilitam que os operadores de rede consigam um eficaz provisionamento e dimensionamento dos recursos computacionais disponíveis, visando um melhor QoS para os usuários finais.

6 Conclusões

O crescente volume de dispositivos conectados na rede, que produzem e consomem dados de maneira colossal, representam desafios relevantes para a próxima geração de redes móveis. Conectar esses dispositivos levará uma enorme quantidade de dados a trafegar pela rede, os quais precisam ser armazenados e/ou processados, com eficiência. Uma abordagem eminente é o uso da computação em nuvem, a qual utiliza DCs localizados no núcleo da rede, para processar e armazenar grandes volumes de dados.

A computação em nuvem foi projetada para ser escalável, entretanto, o aumento descomunal de aplicações e serviços com requisitos de QoS, cada vez mais intolerantes, tornam a nuvem tradicional inadequada para cenários que as envolvem. Desta forma, para lidar com tais dilemas, a ECC foi proposta, levando serviços em nuvem para próximo dos usuários finais, na rede de borda. Assim, com a aproximação de DCs para a borda da rede, de forma ordenada, e com uma distribuição decrescente dos poderes de armazenamento e processamento capaz de atender de forma eficiente os diversos perfis de aplicação, forma-se uma hierarquia de DCs.

A hierarquia de DCs que se forma, tem como objetivo, facilitar o rápido desenvolvimento de sistemas complexos e atender certos requisitos, tais como: alto desempenho e capacidade de gerenciamento, aumento da disponibilidade, entre outros. Desta forma, ela soluciona diversos problemas, como, a sobrecarga dos *links* de *backhaul*, que podem criar gargalos e pontos únicos de falha. A ECC então, é bastante discutida na academia e na industria, sendo a implantação de recursos de DCs para atender aos requisitos mínimos de QoS, um dos pontos mais comentados.

O planejamento e dimensionamento desses recursos deve, então, ser pensado e discutido de forma cautelosa, para atender a demanda procedente da quinta geração de redes móveis. Sendo assim, este trabalho apresentou uma formulação matemática para o dimensionamento e provisionamento de recursos em uma hierarquia de DCs, a qual surgiu como um facilitador para o processamento e/ou armazenamento de aplicações e é uma alternativa flexível para obter recursos computacionais sob demanda para variados serviços.

Os resultados mostraram que, o modelo proposto obteve uma melhor utilização dos recursos computacionais disponíveis na rede e dos *links* de *backhaul*, obtendo um uso 50% maior que o FF. Destaca-se também, que o modelo proposto consegue alocar, em média, 92% das aplicações emergentes, esperadas para o 5G. Em contrapartida, o FF

obtém, uma média, de 36% de aplicações alocadas.

O modelo proposto, é capaz de alocar mais aplicações e realizar o uso mais eficiente dos recursos de rede quando comparado à abordagem FF. A hierarquia de DCs atrelada a um eficiente estratégia de alocação será, então, capaz de atender os mais diversos requisitos de QoS, de forma satisfatória, do conjunto de aplicações que está surgindo. Nesse sentido, enfatizamos a importância da caracterização das novas aplicações para uma melhor utilização dos recursos disponíveis.

6.1 Contribuições da Dissertação

- Realização de uma pesquisa relacionada aos principais conceitos de computação em nuvem, otimização linear e computação de borda, junto com uma revisão bibliográfica dos trabalhos relacionados ao tema.
- A formulação matemática para otimização do uso dos recursos em uma hierarquia de DCs, no qual o modelo proposto possui flexibilidade para abranger outros diversos problemas que rodeiem o tema abordado.
- Aplicação do modelo em um estudo de caso comparativo, envolvendo o cenário de computação em nuvem tradicional e o hierarquicamente distribuído, destacando os pontos fortes da hierarquia de DCs.
- Através dos resultados, foi possível mostrar o aproveitamento da distribuição de DCs hierarquicamente atrelado ao modelo proposto e uma discussão sobre os comportamentos dos diferentes perfis de aplicação.
- A divulgação do trabalho por meio de publicação de artigo em conferência internacional, no qual são apresentados a proposta e os resultados do estudo de caso desta dissertação.

  FOGAROLLI, RAFAEL.; PEREIRA, P. H. A.; CARDOSO, DIEGO. Resource Allocation Optimization for Hierarchical Cloud Data Centers In: 4th International Conference on Cloud Computing Technologies and Applications, 2018, Bruxelas.

6.2 Contribuições Adicionais

Artigos publicados não inclusos neste trabalho.

- ARAUJO, WELTON; FOGAROLLI, RAFAEL; SERUFFO, MARCOS; CARDOSO, DIEGO Deployment of small cells and a transport infrastructure concurrently for next-generation mobile access networks. PLoS One. , v.13, p.e0207330 - , 2018.

- GONCALVES, M. P. S.; LETO, M. B.; FOGAROLLI, RAFAEL.; BARROS, F. J. B.; CARDOSO, D.. Offloading approach for a Two-Level Allocation in H-CRAN Architecture In: 4th International Conference on Cloud Computing Technologies and Applications, 2018, Bruxelas.

- RAMOS DA PAIXAO, ERMINIO AUGUSTO; VIEIRA, RAFAEL FOGAROLLI; ARAUJO, WELTON VASCONCELOS; CARDOSO, DIEGO LISBOA Optimized load balancing by dynamic BBU-RRH mapping in C-RAN architecture In: 2018 Third International Conference on Fog and Mobile Edge Computing (FMEC), 2018, Barcelona.

- FOGAROLLI, RAFAEL.; PAIXAO, E.; SOUZA, D. S.; ARAUJO, W. V.; CARDOSO, D.. Mapeamento BBU-RRH Utilizando Algoritmo Bat In: VIII CONFERÊNCIA NACIONAL EM COMUNICAÇÕES, REDES E SEGURANÇA DA INFORMAÇÃO, 2018, Salvador.

- GONCALVES, M. P. S.; LETO, M. B.; VIEIRA, R. F.; BARROS, F. J. B.; CARDOSO, D.. Abordagem de Offloading para Alocação de Dois Níveis em uma Arquitetura H-CRAN In: VIII CONFERÊNCIA NACIONAL EM COMUNICAÇÕES, REDES E SEGURANÇA DA INFORMAÇÃO, 2018, Salvador.

- PAIXAO, E.; CARDOSO, J.; FOGAROLLI, RAFAEL.; CARDOSO, DIEGO. Análise do Fluxo Multimídia e Desempenho dos Algoritmos de Roteamentos em Redes Vanets V2V/V2I In: I Congresso de Tecnologias e Desenvolvimento da Amazônia, 2017, Cametá.

- FOGAROLLI, RAFAEL.; PAIXAO, E.; CARDOSO, DIEGO. Comportamento da SINR e da Capacidade do Canal com o aumento da Frequência In: I Congresso de Tecnologias e Desenvolvimento da Amazônia, 2017, Cametá.

6.3 Trabalhos Futuros

Como possíveis desdobramentos deste trabalho, compreende-se:

- A realização de novos estudos de caso com cenários heterogêneos, integrando diversas propostas de ECC, tais como, FOG e MEC, possibilitando as aplicações obterem maiores disponibilidades de recursos computacionais.

- Aplicar o modelo proposto em cenários mais realistas através de simulações, acrescentando mobilidade aos usuários. Utilizar uma distribuição de Poisson para o comportamento de chegada das aplicações para verificar o comportamento do sistema com chegadas em instantes diferentes.

- Realizar estudos de caso com topologias diferentes, com distribuições mais aleatórias das posições dos DCs.
- Utilização de técnicas de inteligência computacional, como, heurísticas ou meta-heurísticas para o desenvolvimento eficiente de alocação de recursos, abrindo mão da solução ótima por uma boa, em tempo computacional hábil.
- Aumentar a complexidade do modelo proposto, acrescentando variáveis e parâmetros que foram abstraídos na formulação atual, tais como, modulação do canal, ruídos externos, custo de implantação e uso dos recursos computacionais.

6.4 Dificuldades Encontradas

Os desafios apresentados para conclusão deste trabalho são elencados a seguir:

- O levantamento bibliográfico detalhado sobre os conceitos que permeiam o tema de hierarquia de DCs.
- O estudo da linguagem OPL (*Optimization Programming Language*), necessária para utilização do *software IBM CPLEX Optimization Studio* ®.
- Ausência de informações precisas sobre os variados perfis de aplicações emergentes, determinantes para uma análise satisfatória do comportamento delas na rede.
- A grande quantidade de execuções do modelo proposto e o tempo necessário para realizá-las.

Referências

AGIWAL, M.; ROY, A.; SAXENA, N. Next generation 5g wireless networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 2016. IEEE, v. 18, n. 3, p. 1617–1655, 2016. Citado na página 14.

ALAM, A. B.; ZULKERNINE, M.; HAQUE, A. A reliability-based resource allocation approach for cloud computing. In: IEEE. *2017 IEEE 7th International Symposium on Cloud and Service Computing (SC2)*. [S.l.], 2017. p. 249–252. Citado na página 18.

ALI-YAHIYA, T. *Understanding LTE and its Performance*. [S.l.]: Springer Science & Business Media, 2011. Citado na página 13.

ALNAZIR, M. K. A. M. et al. Performance analysis of cloud computing for distributed data center using cloud-sim. In: IEEE. *2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*. [S.l.], 2017. p. 1–6. Citado na página 21.

AMADEO, M. et al. Information-centric networking for the internet of things: challenges and opportunities. *IEEE Network*, 2016. IEEE, v. 30, n. 2, p. 92–100, 2016. Citado na página 15.

ANDREWS, J. G. et al. What will 5g be? *IEEE Journal on selected areas in communications*, 2014. IEEE, v. 32, n. 6, p. 1065–1082, 2014. Citado na página 14.

ARAL, A.; BRANDIC, I. Quality of service channelling for latency sensitive edge applications. In: IEEE. *2017 IEEE International Conference on Edge Computing (EDGE)*. [S.l.], 2017. p. 166–173. Citado na página 19.

BENGUEDDACH, A.; NIAR, S.; BELDJILALI, B. Online first fit algorithm for modeling the problem of configurable cache architecture. In: IEEE. *2011 International Conference on Microelectronics (ICM)*. [S.l.], 2011. p. 1–6. Citado na página 4.

BORKAR, S.; PANDE, H. Application of 5g next generation network to internet of things. In: IEEE. *International Conference on Internet of Things and Applications (IOTA)*. [S.l.], 2016. p. 443–447. Citado 2 vezes nas páginas 15 e 16.

CAMPEANU, G. A mapping study on microservice architectures of internet of things and cloud computing solutions. In: IEEE. *2018 7th Mediterranean Conference on Embedded Computing (MECO)*. [S.l.], 2018. p. 1–4. Citado na página 1.

CHEN, J. A cloud resource allocation method supporting sudden and urgent demands. In: IEEE. *2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)*. [S.l.], 2018. p. 66–70. Citado na página 18.

CISCO, V. *Cisco Visual Networking Index: Forecast and Methodology 2016–2021.(2017)*. 2017. Citado na página 1.

CORCORAN, P.; DATTA, S. K. Mobile-edge computing and the internet of things for consumers: Extending cloud computing and services to the edge of the network. *IEEE Consumer Electronics Magazine*, 2016. IEEE, v. 5, n. 4, p. 73–74, 2016. Citado na página 2.

DAHLMAN, E.; PARKVALL, S.; SKOLD, J. *5G NR: The next generation wireless access technology*. [S.l.]: Academic Press, 2018. Citado na página 13.

DALLA-COSTA, A. G. et al. Maestro: An nfv orchestrator for wireless environments aware of vnf internal compositions. In: IEEE. *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*. [S.l.], 2017. p. 484–491. Citado na página 21.

DANTZIG, G. *Linear programming and extensions*. [S.l.]: Princeton university press, 2016. Citado na página 11.

DANTZIG, G. B. Maximization of a linear function of variables subject to linear inequalities. *New York*, 1951. 1951. Citado na página 11.

DATTA, P.; SHARMA, B. A survey on iot architectures, protocols, security and smart city based applications. In: IEEE. *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. [S.l.], 2017. p. 1–5. Citado 2 vezes nas páginas 15 e 16.

FORD, R. et al. Provisioning low latency, resilient mobile edge clouds for 5g. In: IEEE. *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. [S.l.], 2017. p. 169–174. Citado na página 19.

GONZALEZ, N. M. et al. Fog computing: Data analytics and cloud distributed processing on the network edges. In: IEEE. *2016 35th International Conference of the Chilean Computer Science Society (SCCC)*. [S.l.], 2016. p. 1–9. Citado na página 20.

GUPTA, A.; JHA, R. K. A survey of 5g network: Architecture and emerging technologies. *IEEE access*, 2015. IEEE, v. 3, p. 1206–1232, 2015. Citado na página 14.

HAYES, B. Cloud computing. *Communications of the ACM*, 2008. ACM, v. 51, n. 7, p. 9–11, 2008. Citado 2 vezes nas páginas 6 e 7.

HEJAZI, H. et al. Survey of platforms for massive iot. In: IEEE. *2018 IEEE International Conference on Future IoT Technologies (Future IoT)*. [S.l.], 2018. p. 1–8. Citado na página 15.

HILLIER, F. S.; LIEBERMAN, G. J. *Introdução à pesquisa operacional*. [S.l.]: McGraw Hill Brasil, 2013. Citado na página 12.

HONG, H.-J. From cloud computing to fog computing: Unleash the power of edge and end devices. In: IEEE. *2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*. [S.l.], 2017. p. 331–334. Citado na página 21.

HOSSAIN, E.; HASAN, M. 5g cellular: key enabling technologies and research challenges. *IEEE Instrumentation & Measurement Magazine*, 2015. IEEE, v. 18, n. 3, p. 11–21, 2015. Citado na página 14.

HU, Y. C. et al. Mobile edge computing—a key technology towards 5g. *ETSI white paper*, 2015. v. 11, n. 11, p. 1–16, 2015. Citado na página 11.

JAMEEL, F. et al. Wireless social networks: A survey of recent advances, applications and challenges. *IEEE Access*, 2018. IEEE, 2018. Citado na página 15.

JANSEN, W.; GRANCE, T. Sp 800-144. guidelines on security and privacy in public cloud computing. 2011. National Institute of Standards & Technology, 2011. Citado na página 8.

JARARWEH, Y. et al. The future of mobile cloud computing: integrating cloudlets and mobile edge computing. In: IEEE. *2016 23rd International Conference on Telecommunications (ICT)*. [S.l.], 2016. p. 1–5. Citado 2 vezes nas páginas 10 e 21.

JIJIN, J. et al. Service load balancing in fog-based 5g radio access networks. In: IEEE. *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. [S.l.], 2017. p. 1–5. Citado na página 19.

KHAN, A. M.; FREITAG, F. On edge cloud service provision with distributed home servers. In: IEEE. *2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*. [S.l.], 2017. p. 223–226. Citado 2 vezes nas páginas 21 e 22.

LEE, I.; LEE, K. The internet of things (iot): Applications, investments, and challenges for enterprises. *Business Horizons*, 2015. Elsevier, v. 58, n. 4, p. 431–440, 2015. Citado na página 15.

LENSTRA, J. K.; KAN, A. R.; SCHRIJVER, A. History of mathematical programming: a collection of personal reminiscences. 1991. CWI, 1991. Citado na página 11.

LÓPEZ-PIRES, F. Many-objective resource allocation in cloud computing datacenters. In: IEEE. *2016 IEEE International Conference on Cloud Engineering Workshop (IC2EW)*. [S.l.], 2016. p. 213–215. Citado na página 17.

MACULAN, N.; FAMPA, M. H. C. Otimização linear. *Editora UNB, 1a edição*, 2006. 2006. Citado 2 vezes nas páginas 11 e 12.

MEHTA, A. et al. How beneficial are intermediate layer data centers in mobile edge networks? In: IEEE. *IEEE International Workshops on Foundations and Applications of Self* Systems*. [S.l.], 2016. p. 222–229. Citado na página 10.

MELL, P.; GRANCE, T. et al. The nist definition of cloud computing. *National institute of standards and technology*, 2009. v. 53, n. 6, p. 50, 2009. Citado na página 6.

NAKANO, M. K.; ALMEIDA, R. de; STEINER, M. T. A. Automotive industry line board optimization through operations research techniques. *IEEE Latin America Transactions*, 2018. IEEE, v. 16, n. 2, p. 585–591, 2018. Citado na página 4.

NGUYEN, T.-D.; HUH, E.-N. Ecsim++: An inet-based simulation tool for modeling and control in edge cloud computing. In: IEEE. *2018 IEEE International Conference on Edge Computing (EDGE)*. [S.l.], 2018. p. 80–86. Citado na página 20.

NUNNA, S. et al. Enabling real-time context-aware collaboration through 5g and mobile edge computing. In: IEEE. *2015 12th international conference on information technology-new generations (ITNG)*. [S.l.], 2015. p. 601–605. Citado na página 10.

PALATTELLA, M. R. et al. Internet of things in the 5g era: Enablers, architecture, and business models. *IEEE Journal on Selected Areas in Communications*, 2016. IEEE, v. 34, n. 3, p. 510–527, 2016. Citado na página 15.

RAHMAN, M. S. et al. A testbed for collecting qos data of cloud-based analytic services. In: IEEE. *2016 IEEE 9th International Conference on Cloud Computing (CLOUD)*. [S.l.], 2016. p. 236–243. Citado na página 18.

RAO, S. S. *Engineering optimization: theory and practice*. [S.l.]: John Wiley & Sons, 2009. Citado 2 vezes nas páginas 11 e 12.

RIMAL, B. P.; CHOI, E.; LUMB, I. A taxonomy and survey of cloud computing systems. In: IEEE. *Fifth International Joint Conference on INC, IMS and IDC, 2009. NCM'09.* [S.l.], 2009. p. 44–51. Citado na página 7.

RUAN, L. et al. The re-expanded cloud: Distributed uplink offloading for mobile edge computing. In: IEEE. *2018 IEEE International Conference on Communications (ICC)*. [S.l.], 2018. p. 1–6. Citado na página 18.

SHAFI, M. et al. 5g: A tutorial overview of standards, trials, challenges, deployment, and practice. *IEEE Journal on Selected Areas in Communications*, 2017. IEEE–Institute of Electrical and Electronics Engineers Inc., v. 35, n. 6, p. 1201–1221, 2017. Citado na página 14.

SILVA, C. N. da et al. Restoration in optical cloud networks with relocation and services differentiation. *Journal of Optical Communications and Networking*, 2016. Optical Society of America, v. 8, n. 2, p. 100–111, 2016. Citado na página 18.

SKALA, K. et al. Scalable distributed computing hierarchy: Cloud, fog and dew computing. *Open Journal of Cloud Computing (OJCC)*, 2015. RonPub, v. 2, n. 1, p. 16–24, 2015. Citado na página 3.

STURZINGER, E.; TORNATORE, M.; MUKHERJEE, B. Application-aware resource provisioning in a heterogeneous internet of things. In: IEEE. *2017 International Conference on Optical Network Design and Modeling (ONDM)*. [S.l.], 2017. p. 1–6. Citado 2 vezes nas páginas 17 e 21.

STURZINGER, E.; TORNATORE, M.; MUKHERJEE, B. Application-aware resource provisioning in a heterogeneous internet of things. In: IEEE. *2017 International Conference on Optical Network Design and Modeling (ONDM),*. [S.l.], 2017. p. 1–6. Citado 2 vezes nas páginas 24 e 25.

VELASCO, L.; RUIZ, M. Flexible fog computing and telecom architecture for 5g networks. In: IEEE. *2018 20th International Conference on Transparent Optical Networks (ICTON)*. [S.l.], 2018. p. 1–4. Citado na página 20.

VOAS, J. Networks of 'things'. *NIST Special Publication*, 2016. v. 800, n. 183, p. 800–183, 2016. Citado na página 15.

WANG, S. et al. A survey on mobile edge networks: Convergence of computing, caching and communications. *IEEE Access*, 2017. IEEE, v. 5, p. 6757–6779, 2017. Citado 4 vezes nas páginas 2, 6, 10 e 11.

ZHANG, J. et al. Joint computation offloading and resource allocation optimization in heterogeneous networks with mobile edge computing. *IEEE Access*, 2018. IEEE, v. 6, p. 19324–19337, 2018. Citado na página 17.

ZHANG, X. et al. Multi-objective resource allocation for mobile edge computing systems. In: IEEE. *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. [S.l.], 2017. p. 1–5. Citado na página 18.

ZHARIKOV, E.; ROLIK, O.; TELENYK, S. A decomposition approach to hierarchical management of cloud data center services. In: IEEE. *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*. [S.l.], 2018. v. 1, p. 42–47. Citado na página 17.