Um Esquema de Multicaminhos com Algoritmos Genéticos para Redes de Centro de Dados

Lyno Henrique Gonçalvez Ferraz, Diogo Menezes Ferrazani Mattos e Otto Carlos Muniz Bandeira Duarte

[1]Grupo de Teleinformática e Automação
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brasil

{lyno,menezes,otto}@gta.ufrj.br

Resumo. *Os centros de dados utilizados para computação em nuvem devem permitir a coexistência de serviços com padrões de tráfego distintos, garantir alta capacidade de transmissão de dados e tolerar falhas de enlaces. As topologias de interconexão dos centros de dados proveem redundância nas conexões físicas, que os mecanismos de encaminhamento utilizam para gerar múltiplos caminhos, e assim melhorar o desempenho de encaminhamento de pacotes. Este artigo propõe um esquema de geração de multicaminhos baseado em algoritmos genéticos que minimiza o tamanho dos caminhos e maximiza diversidade de enlaces usados na rede. Um simulador de eventos discretos foi desenvolvido para a avaliação das técnicas multicaminhos. O simulador modela o comportamento dos fluxos em diversos cenários de centros de dados. O artigo compara o esquema proposto com técnicas de uso de multicaminhos em redes de centro de dados propostos na literatura. Os resultados mostram que a proposta alcança maior taxa média de transmissão de fluxos, mesmo em cenários de alta utilização da rede.*

Abstract. *Data centers used in cloud computing should allocate services with different traffic patterns, provide high data transfer capacity and link fault tolerance. The data center network topologies provide physical connection redundancy, which forwarding mechanisms avail to generate multiple paths. This paper proposes a multipathing scheme based on genetic algorithms to minimize path lengths and maximize link usage diversity. We develop a flow simulator to evaluate the multipathing techniques. The simulations model flow behaviors in different data center scenarios and compare the proposed scheme with multipathing techniques in literature. The results show the proposed scheme highest transmission rates, even in high network utilization scenarios.*

1. Introdução

A geração de dados cresce de forma exponencial e diversas tecnologias tais como Internet das coisas e redes elétricas inteligentes [Guimarães et al., 2013] devem contribuir ainda mais para esse cenário. O armazenamento e o tratamento dessas grandes massas de dados são uma área denominada *Big Data* [Costa et al., 2012] que impõe enormes desafios tecnológicos e, na qual, a computação em nuvem e os centros de dados vão desempenhar papéis fundamentais. O número de aplicações hospedadas nos centros de dados têm aumentado e uma das principais demandas de centros de dados em nuvem é a alta

Este trabalho foi realizado com recursos da CNPq, CAPES, FAPERJ, FINEP e FUNTTEL.

taxa de transmissão de fluxos entre os servidores. Os centros de dados oferecem uma grande capacidade de processamento e armazenamento para aplicações ao aglomerar servidores interconectados. As aplicações são distribuídas nesse aglomerado de servidores, então a rede de comunicação possui papel fundamental para que as aplicações executem de acordo com seus objetivos [Bari et al., 2013].

As topologias das redes de comunicação dos centros de dados são desenvolvidas para prover alta taxa de transmissão agregada, redundância de caminhos e confiabilidade. Para tal, as topologias são formadas com árvores com múltiplas raízes, de modo que ofereçam múltiplos caminhos entre pares de servidores. Por sua vez, mecanismos de encaminhamento normalmente utilizam os caminhos redundantes para tolerância a falhas [Couto et al., 2012]. No cenário de computação em nuvem, diversas aplicações de proprietários diferentes, ou inquilinos, compartilham tanto os servidores quanto a rede dos centros de dados de um provedor de infraestrutura. As cargas de trabalhos de cada inquilino são desconhecidas pelo provedor de infraestrutura, pois cada inquilino executa suas próprias aplicações e protocolos de comunicação. Portanto, o provedor de infraestrutura deve oferecer alta taxa de transmissão sem modificações em protocolos ou *software*. A infraestrutura deve empregar técnicas de uso dos multicaminhos disponíveis.

Este artigo propõe um esquema de criação de multicaminhos baseado em algoritmo genético, cuja função objetivo é minimizar os tamanhos dos caminhos e maximizar a diversidade de uso dos enlaces. Além disso, são modeladas heurísticas para seleção de caminhos já criados. O trabalho também apresenta as demandas de centros de dados em nuvem e analisa algumas técnicas de uso de multicaminhos. O artigo compara o esquema de multicaminhos proposto com as principais técnicas de encaminhamento de tráfego empregadas em centro de dados, como *Spanning Tree Protocol* [Touch e Perlman, 2009], *Equal Cost MultiPath* [Al-Fares et al., 2010], *Smart Path Assignment In Networks* (SPAIN) [Mudigonda et al., 2010].

Um simulador de fluxos de eventos discretos foi desenvolvido para a análise e comparação de desempenho das técnicas de multicaminhos. A simulação de fluxos permite uma maior escala e carga de simulação em relação a simuladores de pacotes. O simulador usa um modelo simplificado de fluxos que disputam a banda disponível nos enlaces dos centros de dados. Os resultados obtidos comprovam que o esquema proposto alcança maior taxa média de transmissão de fluxos, quando comparado com técnicas da literatura, tanto em cenários de comunicação todos-com-todos e todos-com-um.

O artigo está organizado da seguinte forma. A Seção 2 apresenta questões dos centros de dados em nuvem virtualizados. A Seção 3 apresenta o algoritmo genético para a criação de multicaminhos. O simulator desenvolvido é apresentado na Seção 4 e os resultados na Seção 5. Por fim, a Seção 6 conclui o artigo.

2. Sistema de Comunicação em Centros de Dados Virtualizados

Os centros de dados convencionais usam servidores dedicados para executar aplicações específicas, o que resulta na utilização ineficiente dos recursos devido à variação de demanda e da conseqüente ociosidade de recursos. Com o crescimento da computação em nuvem e as tecnologias de virtualização, diversos serviços e aplicações são agregados em um mesmo centro de dados para aumentar a utilização dos recursos físicos e, assim, reduzir os custos de operação e manutenção. Os centros de dados em nuvem virtualizados hospedam inquilinos diversos com suas próprias aplicações, o que aumenta a quantidade e variedade de aplicações que compartilham a infraestrutura do centro de dados [Bari et al., 2013, Mattos e Duarte, 2012].

A infraestrutura dos centros de dados é construída com o objetivo de prover alta capacidade de computação para a execução de aplicações, muitas vezes com topologias altamente redundantes, de modo que sempre haja múltiplos caminhos entre os servidores [Al-Fares et al., 2008, Costa et al., 2012]. Entretanto, uma aplicação é divida e distribuída entre os servidores, de maneira que o tráfego interno de um centro de dados é quatro vezes maior que o tráfego externo [Greenberg et al., 2011].

As aplicações que executam em um centro de dados são variadas, com padrões de tráfegos heterogêneos [Benson et al., 2010]. A maior parte do "tráfego" transmitido entre as aplicações é composta de fluxos com pequena quantidade de dados e muito breves, chamados de fluxos camundongos (*mice flows*). Por outro lado, a maior parte dos "dados transmitidos" está em fluxos com grande quantidade de dados e duradouros, que compõe uma minoria do total de fluxos transmitidos em um centro de dados. Esses fluxos volumosos e duradouros são chamados de fluxos elefantes (*elephant flows*). Os fluxos camundongos e fluxos elefantes possuem requisitos e comportamentos distintos, e a interação causa perdas de pacotes e atrasos no cumprimento de requisições o que impacta no desempenho das aplicações. Os fluxos camundongos são sensíveis à latência, pois normalmente provêm de aplicações do tipo partição/agregação, como *Map/Reduce* usadas em buscas *web*, composição de conteúdos de redes sociais e seleção de propaganda [Alizadeh et al., 2010]. Nesse tipo de aplicação, na fase de partição, uma requisição é dividida em diversas sub-requisições por um servidor de agregação e enviadas a diversos servidores de trabalho. Após a computação em paralelo das requisições, os servidores de trabalho enviam as repostas ao servidor de agregação que compõe a resposta à requisição e, assim, realiza a fase de agregação. Os fluxos desse tipo de aplicação possuem estritos limites de latência para que as repostas sejam apresentadas a tempo para os usuários. As respostas que ultrapassarem o limite de tempo são descartadas, degradando a qualidade da resposta para os usuários. A principal causa de aumento na latência é o congestionamento instantâneo dos enlaces causado pela fase de agregação, que corresponde a uma comunicação do tipo muitos-para-um. Assim, a prevenção desse problema é a priorização de fluxos com prazo curto e baixa utilização de enlaces [Zats et al., 2012].

Diversas propostas criam mecanismos para controlar a banda e o atraso dos fluxos, para garantir a qualidade e prazo das respostas para os usuários de aplicações em centros de dados. A proposta *Data Center TCP* (DCTCP) [Alizadeh et al., 2010] evita congestionamentos e perda de pacotes nos enlaces com um mecanismo fino de controle de congestionamento baseado em notificações explícitas de congestionamento. Assim, os comutadores do centro de dados marcam *bits* de congestionamento nos pacotes de retorno do fluxo quando detectam que as filas de transmissão estão ocupadas e os emissores reduzem a taxa de encaminhamento proporcionalmente à fração de enlaces congestionados.

A proposta *High-bandwidth Ultra-Low Latency* [Alizadeh et al., 2012] garante a latência mínima ao custo de forçar a taxa de transmissão para ser menor que a capacidade do enlace. Para fazer o controle de taxa, os autores utilizam o mecanismo DCTCP, mas marcam os *bits* de congestionamento com outra estratégia. Cada interface de saída possui uma fila fantasma que contabiliza a taxa de saída de pacotes e, quando a taxa de saída ultrapassa um limiar menor que o limite da fila de transmissão, os *bits* de congestionamento são marcados. A taxa de transmissão nunca é máxima, mas a proposta controla a banda dos fluxos antes de ocorrer congestionamento ou perda de pacotes.

O esquema de controle de protocolo *Deadline-Driven Delivery* [Wilson et al., 2011] prioriza os fluxos com menores prazos de entrega através

do controle de taxa. Periodicamente, as aplicações requisitam taxas aos roteadores de acordo com a quantidade de dados restantes nos fluxos e os prazos de entrega. Os roteadores distribuem taxas para as aplicações com um algoritmo guloso e, portanto, as aplicações transmitem os fluxos nas taxas máximas sem violar o prazo de entrega de nenhum fluxo.

Zats *et al.* propuseram *DeTail* [Zats et al., 2012] que é uma abordagem multicamadas para reduzir o tempo máximo de resposta a requisições. Na camada de enlace, os comutadores evitam perdas de pacotes devido à ocupação de filas com o uso de quadros de pausa para controlar a taxa de pacotes recebidos. Na camada de rede, os comutadores escolhem o próximo salto de pacotes com base na ocupação das filas para fazer o balanceamento de carga. O protocolo da camada de transporte é resistente a reordenamento para receber os pacotes vindos de vários caminhos e controla a taxa de transmissão com notificações de congestionamento de comutadores baseado na ocupação das filas. As aplicações especificam as prioridades para diferenciar fluxos sensíveis à latência.

As propostas até aqui mencionadas não só modificam a infraestrutura para melhorar o desempenho da rede, mas também modificam os servidores finais que devem interagir com a infraestrutura. Logo, essas propostas não são adequadas para centros de dados em nuvem virtualizados com multi-inquilinos, pois os inquilinos possuem protocolos próprios e diversificados. Além disso, mesmo padronizando os protocolos, há riscos de segurança na quebra de isolamento entre inquilinos, pois protocolos de aplicações comunicam-se diretamente com dispositivos da infraestrutura para reservar recursos.

Ao mesmo tempo em que se procura diminuir as perdas devido ao congestionamento instantâneo de enlaces, também se deve atender às características diferentes dos fluxos. Fluxos elefantes demandam alta capacidade de transmissão de dados e são resistentes à latência, pois transferem uma grande quantidade de dados e não têm os mesmos requisitos de entrega de fluxos camundongo. Assim, esses fluxos devem ser organizados de maneira a aproveitar o máximo possível da capacidade de transmissão dos centros de dados. Além disso, fluxos elefantes utilizam toda a capacidade de transmissão de um enlace e, assim, causam congestionamentos que afetam os fluxos camundongos. Diversas propostas aproveitam os múltiplos caminhos que as redes de centros de dados possuem para aproveitar o máximo da capacidade de transmissão dos enlaces. Uma forma de aproveitar os multicaminhos é através de *Equal Cost MultiPath* (ECMP). O protocolo de rede calcula múltiplos menores caminhos de mesmo custo e realiza um *hash* dos cabeçalhos dos pacotes para escolher o caminho no qual transmitir o fluxo. Assim, é esperado que os fluxos sejam distribuídos aleatoriamente nos múltiplos caminhos, ou seja, caminhos com custo igual. Diversos protocolos usam essa técnica de encaminhamento, como *Transparent Interconnection of Lots of Links* [Touch e Perlman, 2009] e 802.1aq *Shortest Path Bridging* [Allan et al., 2010]. O *Valiant Load Balancing* [Greenberg et al., 2011] funciona de maneira semelhante, mas a escolha de caminho é realizada através da escolha aleatória de um comutador intermediário.

A proposta *Smart Path Assignment In Networks* (SPAIN) [Mudigonda et al., 2010] explora a diversidade de caminhos das topologias de centros de dados para aumentar a vazão dos fluxos e a confiabilidade da rede. O SPAIN usa um algoritmo *offline* para configurar as árvores de VLANs (*Virtual Local Area Network*), de modo em que as árvores são construídas com base em menores caminhos e menor uso de enlaces. Outro algoritmo *online* executado nos servidores verifica as árvores e os servidores conectados e, então seleciona aleatoriamente uma árvore para ser usada em um fluxo.

O *MultiPath TCP* (MPTCP) [Raiciu et al., 2011] subdivide um fluxo TCP em diversos subfluxos, de maneira que cada um possua seu próprio controle de congestionamento. Cada um dos subfluxos é transmitido em um caminho diferente e, assim, cada subfluxo transmite na taxa máxima de cada caminho. Os multicaminhos utilizados pelo MPTCP podem ser definidos através de mecanismos como o SPAIN.

A proposta Hedera [Al-Fares et al., 2010] detecta os fluxos elefantes e os escalona nos caminhos dos centros de dados. O Hedera usa o controlador centralizado de redes NOX [Gude et al., 2008] comutadores programáveis OpenFlow [McKeown et al., 2008] para detectar os fluxos com alta taxa de transmissão de dados e tempo de vida. O controlador periodicamente migra os fluxos elefantes para outros caminhos baseado em um algoritmo de otimização de arrefecimento simulado (*simulated annealing*) que otimiza a taxa de transmissão dos fluxos. Essas propostas abordam o problema de organização dos fluxos nos enlaces da rede de centro de dados e todas, exceto MPTCP, são adequadas para centros de dados em nuvem virtualizados, pois não modificam os protocolos de inquilinos e podem ser utilizados por provedores de infraestrutura. Entretanto, a seleção de caminhos aleatória não leva em conta a utilização da rede e os tamanhos dos fluxos, o que causa colisões de caminhos que sobrecarregam enlaces e degradam o desempenho.

Este artigo foca em técnicas multicaminhos e propõe um esquema baseado na otimização da geração de múltiplos caminhos com algoritmos genéticos. A abordagem multicaminhos é mais adequada para o cenário de centros de dados em nuvem virtualizados do ponto de vista de um provedor de infraestrutura, pois não interferem nas aplicações dos inquilinos. Logo, as técnicas de multicaminhos consideradas não devem modificar protocolos de servidores finais, mas somente a infraestrutura de rede. A utilização dos multicaminhos é realizada em duas fases: a Configuração de Multicaminhos e a Seleção de Multicaminhos. Na fase Configuração de Multicaminhos, um algoritmo calcula os caminhos a serem configurados e, em seguida, configura os dispositivos de rede com diversos caminhos. Normalmente a configuração de caminhos é realizada de maneira *offline* ou quando ocorrem mudanças na topologia da rede. A fase de Seleção de Multicaminhos é *online* e ocorre constantemente enquanto a rede está em operação. Nessa fase, os dispositivos de rede usam algoritmos para selecionar qual dos multicaminhos configurados é utilizado para um fluxo.

3. O Esquema Proposto de Multicaminhos com Algoritmo Genético

Para aproveitar a redundância de caminhos na topologia dos centros de dados, este artigo propõe um esquema de Configuração de multicaminhos com algoritmos genéticos e de Seleção com heurísticas baseadas no uso dos caminhos. O esquema gera diversas árvores independentes para interconectar os comutadores de topo de *rack*, de maneira a minimizar tanto as distâncias entre comutadores de topo *rack*, quanto maximizar a diversidade de uso de enlaces na rede. As árvores são configuradas nos comutadores para o encaminhamento dos pacotes, com VLANs ou com controladores centralizados. Em cada árvore existe somente um caminho entre cada par de comutadores topo de *rack*. Somente uma árvore é selecionada para cada fluxo, portanto, o fluxo trafega por um único caminho o que não afeta a ordenação dos pacotes no receptor. A seleção das árvores a serem utilizadas por um fluxo pode ser determinada por diferentes heurísticas baseadas no uso de caminhos e de enlaces. A técnica SPAIN [Mudigonda et al., 2010] utiliza abordagem semelhante para explorar multicaminhos em centros de dados.

As árvores devem explorar ao máximo a diversidade de enlaces do centro de dados. Assim, o uso de diversas árvores aumenta a taxa de transmissão agregada, balanceia

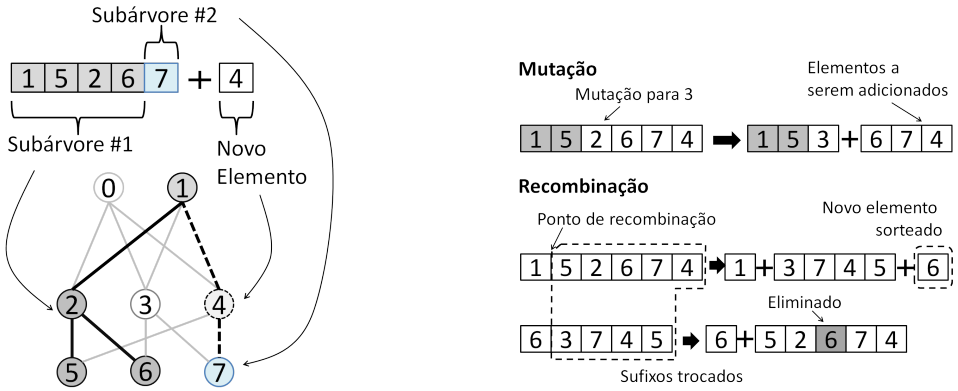
a carga entre os enlaces, diminui o impacto de falhas nos enlaces e previne perdas em rajadas devido a enlaces sobrecarregados. Para configurar as árvores, deve-se obter a topologia de rede do centro de dados para executar o algoritmo de geração de árvores. Após a geração das árvores, configuram-se as árvores nos comutadores. Esse processo é realizado *offline* antes do centro de dados entrar em operação e, portanto, não acarreta em atrasos de configuração e escalonamento de caminhos. Após a configuração das árvore, a seleção de caminhos é realizada por dispositivos inseridos entre servidores e comutadores topo de *rack* ou nos servidores.

Esta abordagem requer modificações mínimas na rede de comunicações do centro de dados. Além disso, essa abordagem não exige nenhuma modificação nas máquinas virtuais de inquilinos, logo é adequada para centros de dados em nuvem virtualizados. Neste artigo, é considerado o caso de uso de VLANs para definir uma árvore, pois exige apenas características de comutadores de prateleira. Para garantir a operação e alcançabilidade na ocorrência de falhas de enlaces das árvores de VLANs, todos comutadores da rede executam o *Spanning Tree Protocol* (STP) e configuram uma árvore de cobertura. Essa árvore de cobertura só é utilizada para o encaminhamento de pacotes que não pertençam a uma VLAN conhecida e ativa.

Algoritmo Genético

A representação de um indivíduo é uma árvore definida pelo arranjo sem repetição de identificadores de comutadores. Para gerar a árvore conexa e sem laços é realizado um processo que interconecta os comutadores do arranjo, que adiciona novos comutadores até conectar todos comutadores topo de *rack*. Uma nova árvore é gerada da seguinte maneira: sorteia-se um comutador e ele é adicionado a uma subárvore. Em seguida, sorteia-se outro comutador. Caso o comutador sorteado possua conexão direta com comutadores já sorteados, o comutador é adicionado na mesma subárvore, assim como todas as subárvores que o comutador possui conexão direta. Os enlaces considerados são aqueles que o comutador usa para conectar-se às árvores. Se o comutador não tem conexão direta com nenhum comutador de outra subárvore, o comutador é adicionado em uma nova subárvore. Esse processo é repetido até todos comutadores topo de *rack* estarem conectados na mesma subárvore. No final, os comutadores que não estão no caminho entre todos comutadores topo de *rack* são removidos da árvore. A Figura 1(a) apresenta um exemplo de genótipo de um indivíduo do algoritmo genético e a árvore formada.

As operações de mutação e recombinação são realizadas de maneira especial para manter a árvore conexa, sem laços. A mutação escolhe e muda aleatoriamente um dos comutadores do genótipo e, caso a mutação separe a árvore em subárvores, são adicionados novos comutadores. Na recombinação, sorteia-se uma posição para manter os comutadores no arranjo de dois genótipos e, os outros comutadores do arranjo são enviados para o outro genótipo. Os comutadores são adicionados sequencialmente e, caso a operação separe a árvore em subárvores, são adicionados novos comutadores até obter árvores conexas. A Figura 1(b) mostra as operações nos genótipos. O fenótipo, ou a avaliação de uma árvore, é representado por duas funções objetivo. A primeira calcula a distância média entre os comutadores topo de *rack* e a segunda calcula a soma do inverso de vezes que um enlace é utilizado em cada árvore. Assim, fenótipos melhores são os de árvores com diâmetros menores e que possuam enlaces menos utilizados. A comparação de fenótipos considera um melhor que outro caso uma função objetivo seja maior e a outra função seja maior ou igual.



(a) Representação de um indivíduo.

(b) Operações sobre o genótipo.

Figura 1. Algoritmo genético para a geração de árvores: a) representação de um indivíduo como um permutação de comutadores; b) operações de mutação e recombinação sobre o genótipo.

O algoritmo é inicializado com certo número de indivíduos e em seguida o algoritmo entra em um laço com um número máximo de gerações. Em cada interação do laço, são sorteados pares de indivíduos proporcionalmente à qualidade do fenótipo. Os pares são recombinados para gerar novos pares de indivíduos até se obter o dobro da população. Em seguida os indivíduos são mutados e, para a próxima geração, sobrevivem apenas os indivíduos de melhor fenótipo dentre todos os indivíduos calculados.

O Procedimento de Seleção de Multicaminhos

Para realizar a seleção de multicaminhos são definidas as seguintes heurísticas:

- seleção aleatória: o caminho é selecionado com probabilidade uniforme;
- seleção de caminhos menos utilizados: cada vez que o caminho é utilizado, diminui a probabilidade de ser selecionado;
- seleção de caminhos com enlaces menos utilizados: cada vez que um enlace de um caminho é utilizado por um fluxo, diminui a probabilidade dos caminhos que usam o enlace serem utilizados.

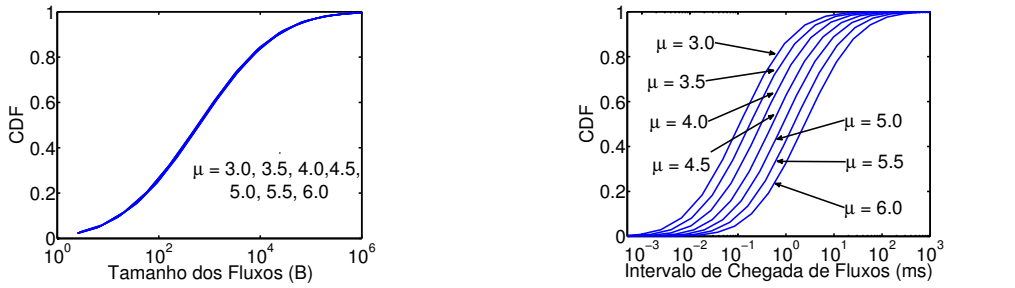
Deve ser ressaltado que os mecanismos que optam pela estratégia de selecionar os caminhos e enlaces menos utilizados precisam obter dados de um banco de informações do uso de caminhos e enlaces, o que causa atraso na seleção de caminhos de novos fluxos.

4. O Simulador de Fluxos de Eventos Discretos Desenvolvido

Para avaliar as técnicas de multicaminhos foi projetado e desenvolvido um simulador de eventos discretos que modela a transmissão de dados como um fluxo de dados. Como o objetivo deste trabalho é a avaliação de desempenho de estratégias de multicaminhos, um modelo de simulação de fluxos permite uma maior escalabilidade de simulação em comparação a modelos de simulação de pacotes como o provido pelo simulador NS3 [ns3, 2006], pois abstrai os procedimentos dos protocolos. Assim, o simulador de fluxos em centro de dados proposto cria uma fila de eventos ordenada pelo instante dos eventos. A cada passo, um novo evento é tratado e o tempo da simulação é atualizado. Eventos podem adicionar novos eventos na fila que são, na maioria, chegada e saída de fluxos. A simulação pára quando não existem mais eventos ou o tempo limite é atingido.

O Modelo: Um fluxo é definido pela tupla (origem, destino, tamanho em *bytes*, taxa de transmissão atual, quantidade *bytes* já transferidos). O modelo do simulador assume

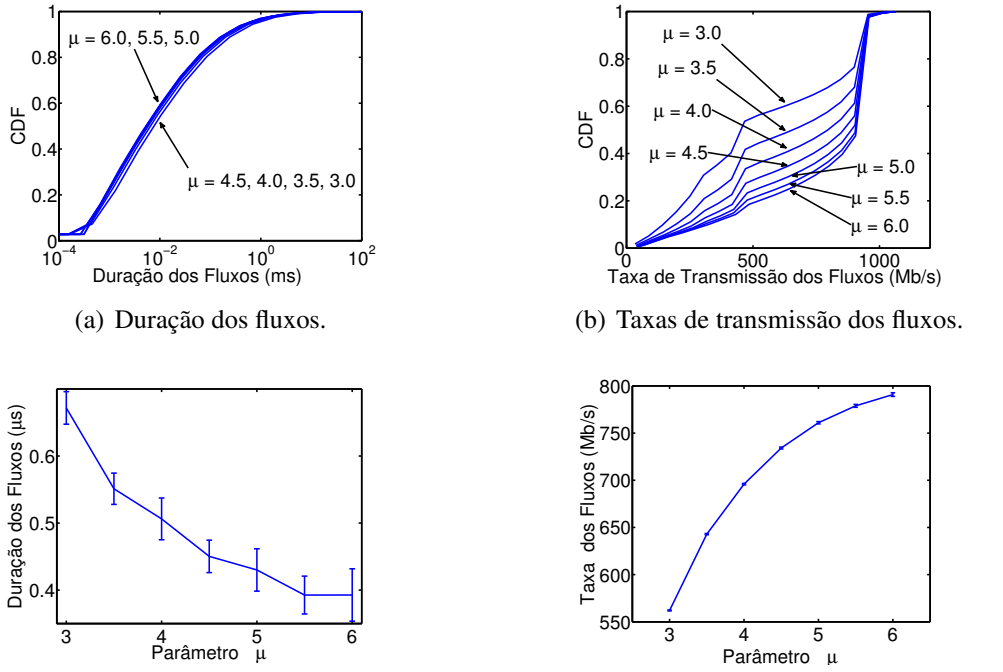
que a taxa de transmissão dos fluxos é máxima taxa obtida em regime permanente e só é modificada com a chegada ou saída de outro fluxo. Assim, a taxa de transmissão é calculada como a divisão igual de banda do enlace entre os fluxos. Caso um fluxo já tenha a taxa definida e ocupe menos que a banda máxima de um enlace, o restante da banda do enlace é dividido igualitariamente entre os fluxos restantes. Além disso, os fluxos são considerados em uma direção e os fluxos correspondentes aos pacotes de retorno dos fluxos são desprezados. Dessa maneira, o modelo é otimista e não considera o início lento do TCP, as perdas de pacote e a banda utilizada pelo fluxo de retorno. O simulador considera os cabeçalhos dos protocolos TCP/IP e da camada de enlace Ethernet e que todos os pacotes são transmitidos com o tamanho máximo permitido, exceto o último pacote de cada fluxo que contém somente os *bytes* restantes do fluxo.



(a) CDF dos tamanhos dos fluxos. (b) CDF dos intervalos de chegada de fluxos.

Figura 2. Tráfego utilizado nas simulações para modelar o intervalo de chegada de fluxos medidos em [Benson et al., 2010]: a) tamanho dos fluxos com distribuição lognormal ($\mu = 7, \sigma = 2.8$); b) intervalos de chegada de fluxos com distribuição lognormal ($\mu = 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, \sigma = 2$).

Os Parâmetros de Simulação: O simulador pode ser configurado com diversos parâmetros que permitem avaliar diferentes aspectos do desempenho dos centros de dados. A topologia do centro de dados indica a diversidade de caminhos e, portanto, avalia a eficácia e desempenho das técnicas de multicaminhos. O número de nós da topologia altera a escala das simulações. A configuração de comportamentos dos fluxos pode ser alterada para variar a carga de trabalho da simulação. O tamanho dos fluxos indica quanto tempo o fluxo ocupa os enlaces do caminho. Quanto maior o tamanho do fluxo em *bytes*, maior o tempo para transmiti-lo por completo, maior a probabilidade de fluxos novos usarem os mesmos enlaces, diminuir a taxa de transmissão e aumentar ainda mais o tempo de transmissão do fluxo. Nas simulações deste artigo, considerou-se uma distribuição lognormal para gerar os tamanhos dos fluxos. Os fluxos são gerados aleatoriamente com distribuição lognormal de média $\mu = 7$, e desvio padrão $\sigma = 2.8$, de modo que a função de distribuição cumulativa tenha os seguintes valores $F(x) = \{\approx 0.5 | x = 1000, \approx 0.95 | x = 100000\}$ de acordo com as medidas empíricas apresentadas por Benson *et al.* [Benson et al., 2010]. De maneira semelhante, os intervalos de chegada de novos fluxos também alteram a carga de trabalho dos centros de dados. Quanto menor o intervalo de chegada, mais fluxos disputam a banda dos enlaces, logo as taxas dos fluxos diminuem e o tempo de transmissão aumenta. As simulações desse artigo usam uma distribuição lognormal para os intervalos de chegada de fluxos com desvio padrão $\sigma = 2$ e média, μ, variando para aumentar a carga de $3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0$. Os valores do parâmetro μ foram escolhidos para modelar o intervalo de chegada de fluxos medidos em [Benson et al., 2010]. Ambos os modelos utilizados nas simulações são apresentados na Figura 2.



(a) Duração dos fluxos.

(b) Taxas de transmissão dos fluxos.

(c) Duração média dos fluxos.

(d) Taxas média de transmissão dos fluxos.

Figura 3. Teste de sanidade do simulador. Duração e taxa de transmissão para diferentes médias μ da distribuição de intervalo de cegada de fluxos. Com a diminuição de μ, mais fluxos disputam a banda do enlace, diminuindo a taxa de transmissão e aumentando a duração dos fluxos.

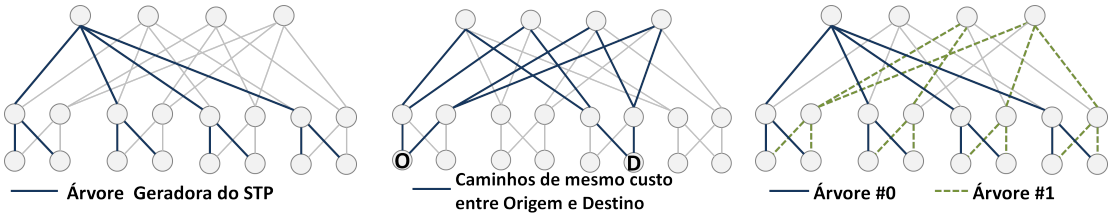
Outra característica importante dos comportamentos dos fluxos é o conjunto de destinos dos fluxos. Se os destinos são escolhidos uniformemente entre todos os nós do centro de dados, é esperado que toda a carga de trabalho seja distribuída uniformemente entre os destinos, apesar das cargas geradas aleatoriamente. Esse cenário utiliza intensamente todos caminhos disponíveis. Por sua vez, caso o tráfego seja concentrado em um ou poucos nós de destino, os caminhos até os destinos são utilizados intensamente enquanto caminhos para outros nós são menos utilizados. Logo, caminhos alternativos que usem enlaces de caminhos para outros nós se tornam uma boa escolha para os fluxos.

Foi realizado um teste de sanidade com uma topologia com dois nós interconectados. Esse teste mostra o impacto do aumento de carga por diminuição do intervalo de chegada de fluxos. Os resultados do teste de sanidade são apresentados na Figura 3. O parâmetro variado foi a média de μ da distribuição lognormal dos intervalo de chegada de fluxos. A Figura 2(a) mostra a função distribuição acumulada (*Cumulative Distribution Function* - CDF) dos tamanhos dos fluxos. A Figura 2(b) mostra a CDF dos intervalos de chegada de fluxos e que a diminuição de μ reduz o valor de intervalo de chegada dos fluxos em média. As Figuras 3(a) e 3(c) mostram a CDF e média das durações dos fluxos e as Figuras 3(b) e 3(d) mostram a CDF e as médias das taxas de transmissão dos fluxos para os diferentes valores de μs. É possível perceber que com a diminuição do μ, a duração dos fluxos tende a valores maiores como apresentado nas Figuras 3(a) e 3(c). A diminuição do μ aumenta o número de fluxos que compartilham o enlace, portanto as taxas de transmissão obtidas são menores como mostram as Figuras 3(b) e 3(d). O restante dos gráficos são apresentados na forma de média das durações dos fluxos e média das taxas de transmissão dos fluxos.

5. Simulações e Resultados

Este artigo avalia as técnicas de multicaminhos considerando as fases de Configuração e Seleção de Multicaminhos. Assume-se que não há falhas em enlaces, nem a reconfiguração de multicaminhos. A topologia utilizada foi a fat-tree [Al-Fares et al., 2008] com comutadores de quatro portas, que oferece quatro caminhos distintos de mesmo custo entre pares de servidores. Essa topologia beneficia a técnica *Equal Cost MultiPath* (ECMP) pela existência de múltiplos caminhos distintos de menor custo, dessa maneira esse é um cenário justo para de comparação. As cargas de trabalho são definidas aleatoriamente por distribuições lognormal, tamanho dos fluxos com distribuição lognormal ($\mu = \mathbf{7}, \sigma = \mathbf{2.8}$), intervalos de chegada de fluxos nos comutadores topo de *rack* com distribuição lognormal ($\mu = \mathbf{3.0}, \mathbf{3.5}, \mathbf{4.0}, \mathbf{4.5}, \mathbf{5.0}, \mathbf{5.5}, \mathbf{6.0}, \sigma = \mathbf{2}$). O simulador executa a simulação até 1000 segundos. Os resultados são apresentados com intervalo de confiança de 95%. Os gráficos são apresentados na forma de média dos valores obtidos na simulação.

A seguir são apresentadas algumas técnicas que foram modeladas, implementadas e comparadas com o esquema proposto de multicaminhos com algoritmo genético.



(a) STP. (b) ECMP. (c) SPAIN.

Figura 4. Multicaminhos na topologia *fattree* criados pelas técnicas a) STP: uma árvore com um caminho único entre pares; b) ECMP: diversos caminhos de mesmo custo entre pares; c) SPAIN: diversas árvores disjuntas.

O Modelo do Spanning Tree Protocol

O *Spanning Tree Protocol* (STP) calcula uma árvore de cobertura entre comutadores, De maneira que os comutadores só aprendem caminhos sobre a árvore. Desse modo, todos os caminhos entre pares de servidores compartilham os enlaces da árvore de cobertura. A Configuração de Multicaminhos dessa técnica considera um único caminho sobre a árvore de cobertura e, portanto, a Seleção de Multicaminhos usa o único caminho disponível. As simulações desse artigo consideram uma árvore de cobertura mínima.

O Modelo do Equal Cost MultiPath

A técnica *Equal Cost MultiPath* (ECMP) calcula todos os caminhos para um destino que têm custo mínimo na fase de Configuração. Neste artigo, o cálculo é realizado com o algoritmo Dijkstra modificado. A Figura 4(b) apresenta os múltiplos caminhos de mesmo custo para a topologia *fattree*. Na fase de Seleção, o elemento de rede aplica uma função *hash* em campos do cabeçalhos dos pacotes para indicar qual próximo salto usar e, assim, um fluxo segue por um único caminho. As simulações utilizam a função *hash* `md5` sobre os identificadores da origem e destino para selecionar o caminhos de um fluxo.

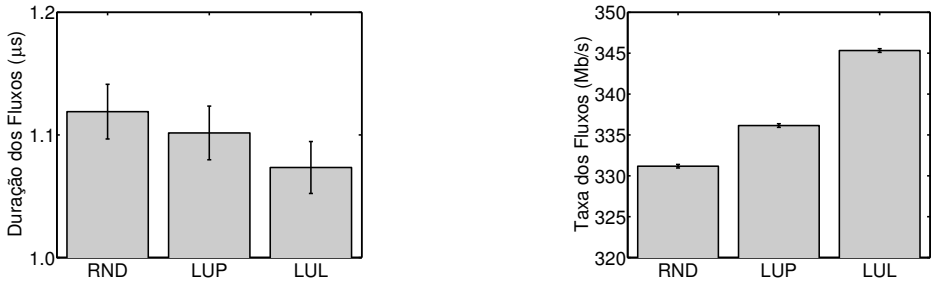
O Modelo do Smart Path Assignment In Networks

Os mecanismos e algoritmos criados pelo *Smart Path Assignment In Networks* (SPAIN) configuram VLANs nos comutadores para cada árvore e, durante a operação da

rede, servidores selecionam uma árvore para um fluxo e marcam a etiqueta de VLAN da árvore. A fase de Configuração calcula de maneira *offline* as múltiplas árvores, que são criadas com dois algoritmos, um forma conjuntos de caminhos distintos entre pares, e outro usa algoritmo guloso para agregar os caminhos entre pares para formar as árvores. Assim, obtém-se um conjunto de árvores com menores caminhos entre pares com enlaces disjuntos. A Figura 4(c) apresenta exemplos de árvores formadas pelo SPAIN. A seleção de caminhos é realizada com um mecanismo que executa nos servidores. O mecanismo consulta uma base de dados de árvores e verifica a disponibilidade dos caminhos. Ao enviar um fluxo, o mecanismo escolhe uniformemente um dos caminhos ativos e marca a etiqueta de VLAN em todos os pacotes desse fluxo.

5.1. Resultados de Simulação com Tráfego Todos-para-Todos

Avalia-se o comportamento dos centros de dados com tráfego de todos os comutadores topo de *rack* para todos outros comutadores topo de *rack*. Como os destinos são escolhidos uniformemente, é esperado que a carga de trabalho seja distribuída uniformemente entre os caminhos com alta ocupação dos enlaces.

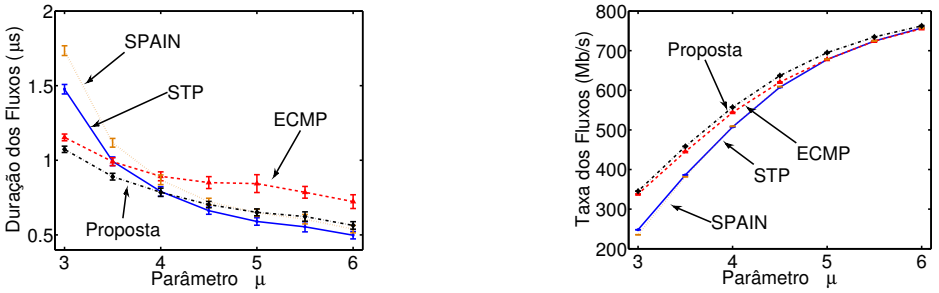


(a) Duração dos fluxos. (b) Taxa de transmissão dos fluxos.

Figura 5. Avaliação das Heurísticas de Seleção dos fluxos para o esquema de multicaminhos com algoritmos genéticos. As heurísticas comparadas são seleção aleatória (*RND*), seleção dos caminhos menos utilizados (*LUP*) e seleção dos caminhos com enlaces menos utilizados (*LUL*).

Avaliação das Heurísticas de Seleção de Multicaminhos

Para escolher a heurística para seleção de multicaminhos do algoritmo genético, comparam-se os seguintes tipos de seleção: seleção aleatória (*Random - RND*), seleção dos caminhos menos utilizados (*Least Used Path - LUP*) e seleção dos caminhos com enlaces menos utilizados (*Least Used Link - LUL*). As heurísticas de seleção de multicaminhos foram comparadas considerando a configuração de multicaminhos com algoritmos genéticos. Nessa simulação, somente uma alta carga de trabalho foi utilizada com intervalos de chegada de fluxos com distribuição lognormal($\mu = \mathbf{3.0}, \sigma = \mathbf{2}$). A Figura 5 apresenta os resultados da simulação para as diferentes heurísticas. A Figura 5(a) mostra que as durações dos fluxos tem valores de aproximadamente 1.1μs, com pouca variação para cada tipo de seleção de caminhos. A Figura 5(b) mostra que apesar da duração dos fluxos serem próximas, as taxas diferem com 331Mb/s para a seleção aleatória (*RND*), 336Mb/s para a seleção dos caminhos menos utilizados (*LUP*) e 345Mb/s para a seleção dos caminhos com enlaces menos utilizados (*LUL*). A seleção dos caminhos com enlaces menos utilizados (*LUL*) diminui a probabilidade de uso de um caminho cujos enlaces já foram escolhidos por muitos fluxos, logo a taxa de transmissão média é maior. O algoritmo genético considerado no restante do artigo é o LUL.



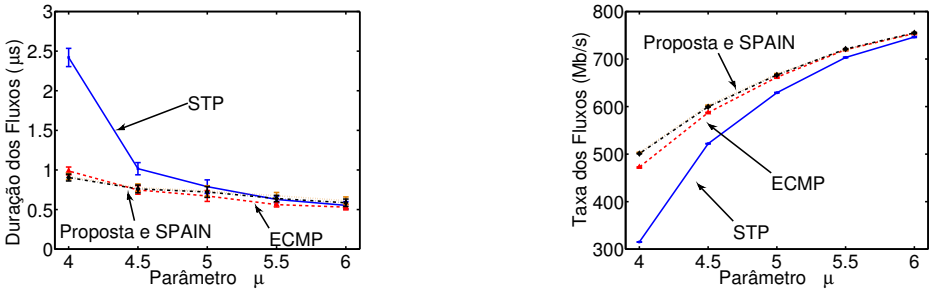
(a) Duração dos fluxos. (b) Taxa de transmissão dos fluxos.

Figura 6. Avaliação de desempenho da Fase de Configuração dos fluxos com tráfego todos-para-todos do esquema proposto de árvores com algoritmos genético e comparação com as técnicas *Spanning Tree Protocol* (STP), *Equal Cost MultiPath* (ECMP), *Smart Path Assignment In Networks* (SPAIN).

O desempenho do esquema proposto de multicaminhos com algoritmos genéticos é avaliado e comparado com as técnicas STP, ECMP e SPAIN, da fase de configuração de multicaminhos. A distribuição dos tamanhos dos fluxos é a mesma da simulação da avaliação das heurísticas de seleção de multicaminhos e o intervalo de chegada de fluxos possui distribuição lognormal com desvio padrão $\sigma = 2$ e com a média μ variando de 3.0 a 6.0. A Figura 6 apresenta os bons resultados da proposta deste artigo para a configuração de multicaminhos. Observa-se que para pequenos valores de μ, a média da duração dos fluxos é maior e a taxa de transmissão é menor, de modo que as técnicas STP e SPAIN apresentam os piores valores. Como no STP existe somente uma mesma árvore em que ocorre a comunicação, todos os fluxos compartilham os mesmos enlaces, o que limita as taxas de transmissão obtidas. De maneira semelhante, a técnica SPAIN cria árvores através de um algoritmo guloso sem considerar caminhos cujos enlaces são menos utilizados nas árvores, então as árvores compartilham muitos enlaces. Além disso, a seleção de caminhos aleatória do SPAIN não prioriza enlaces pouco utilizados, assim o tráfego entre todos os comutadores topo de *rack* ocupa por muito tempo poucos enlaces, o que limita as taxas de transmissão obtidas. Como a topologia *fattree* disponibiliza quatro caminhos diferentes de mesmo custo, a técnica ECMP tem sucesso ao distribuir o tráfego entre os enlaces e atinge altas taxa de transmissão. Entretanto, a seleção de caminhos do ECMP por funções *hash* causa colisões na seleção de caminhos para cada fluxo, aumentando a duração média dos fluxos. O esquema proposto possui menores durações médias de fluxos e maiores taxas médias de transmissão, com ganhos em cima do ECMP que variam de 7,0 a 17,6 Mb/s. O esquema cria árvores com algoritmo genético considerando a utilização de enlaces nas outras árvores, portanto os fluxos usam caminhos distintos. Além disso, a heurística de seleção de caminhos prioriza caminhos com enlaces menos utilizados, o que balanceia o uso de enlaces na rede.

5.2. Avaliação e Comparação de Configuração de Multicaminhos com Tráfego Todos-para-Um

No cenário todos-para-um, o tráfego é concentrado em um comutador de destino que corresponde a fase de agregação de aplicações partição/agregação como *Map/Reduce*. Os comutadores topo de *rack* transmitem diversos fluxos de diferentes tamanhos para um único comutador topo de *rack*. A Figura 7 mostra os resultados da simulação nos quais o esquema proposto sempre apresenta menores durações médias e maiores taxas médias de



(a) Duração dos fluxos.

(b) Taxa de transmissão dos fluxos.

Figura 7. Avaliação de desempenho com tráfego todos-para-um do esquema proposto de configuração de árvores com algoritmos genético e comparação com as técnicas *Spanning Tree Protocol* (STP), *Equal Cost MultiPath* (ECMP), *Smart Path Assignment In Networks* (SPAIN).

transmissão. No STP, todos os fluxos compartilham o enlace diretamente conectado ao comutador de destino e, portanto, cria-se um gargalo. Por outro lado, as técnicas SPAIN, ECMP e também o esquema proposto de algoritmo genético na *fattree* utilizam caminhos diferentes para cada fluxo. ECMP perde em desempenho para o esquema proposto devido às colisões na seleção de enlaces causada pelo *hash* e com $\mu = 4$ o esquema proposto tem vazão 30 Mb/s maior. O esquema proposto com algoritmos genéticos e a técnica SPAIN consideram caminhos alternativos pouco utilizados que podem ser maiores que os caminhos mínimos da rede e, assim, atingem as menores durações médias e as maiores taxas médias de transmissão.

6. Conclusão

As redes de centros de dados em nuvem virtualizados demandam novas técnicas que suportem os fluxos gerados pela grande quantidade e diversidade de aplicações dos múltiplos inquilinos. Os provedores de infraestrutura dos centros de dados em nuvem devem utilizar mecanismos que não interfiram na autonomia e prejudiquem o isolamento dos inquilinos. Esse artigo propõe um esquema de multicaminhos com algoritmos genéticos para que provedores de infraestrutura realizem o encaminhamento eficiente de fluxos. Um simulador de fluxos de eventos discretos foi desenvolvido para a avaliação e comparação da proposta com trabalhos da literatura. A proposta distribui o tráfego em enlaces da rede alcançando maiores taxas de transmissão de fluxos, mesmo em cenários de tráfego todos-para-todos e todos-para-um. O principal ganho da proposta deve-se ao uso de heurísticas de seleção de caminhos que consideram enlaces menos usados ao definir um novo fluxo. A etapa de seleção proposta diferencia-se das demais abordagens de multicaminhos, pois considera a quantidade de fluxos já existente nos enlaces ao definir o novo fluxo. A etapa de configuração de multicaminhos garante também o melhor desempenho da proposta à medida que calcula árvores otimizadas e disjuntas de cobertura da rede. Outro ponto importante da proposta é que se baseia somente na definição de regras de encaminhamento na camada de enlace e, portanto, não depende de mudanças em sistemas operacionais ou protocolos. A principal vantagem é as estações que acessam a rede não precisam ser alteradas para adoção da proposta, mas somente há a configuração de comutadores no núcleo da rede. Assim, a proposta é adequada para o cenário de múltiplos inquilinos compartilhando a rede com multicaminhos de um centro de dados para nuvem.

Como trabalhos futuros serão avaliadas diferentes topologias e estender o simulador para considerar os atrasos dos enlaces e o tamanho de *buffers* dos comutadores para estudar os impactos de uso de multicaminhos em fluxos camundongos.

Referências

[Al-Fares et al., 2008] Al-Fares, M., Loukissas, A. e Vahdat, A. (2008). A scalable, commodity data center network architecture. Em *Proceedings of the ACM SIGCOMM 2008 conference on Data communication*, SIGCOMM '08, p. 63–74. ACM.

[Al-Fares et al., 2010] Al-Fares, M., Radhakrishnan, S., Raghavan, B., Huang, N. e Vahdat, A. (2010). Hedera: Dynamic flow scheduling for data center networks. Em *Proceedings of the 7th USENIX NSDI conference*, p. 19–19. USENIX Association.

[Alizadeh et al., 2010] Alizadeh, M., Greenberg, A., Maltz, D. A., Padhye, J., Patel, P., Prabhakar, B., Sengupta, S. e Sridharan, M. (2010). Data center TCP (DCTCP). Em *Proceedings of the ACM SIGCOMM 2010 conference*, SIGCOMM '10, p. 63–74, New York, NY, USA. ACM.

[Alizadeh et al., 2012] Alizadeh, M., Kabbani, A., Edsall, T., Prabhakar, B., Vahdat, A. e Yasuda, M. (2012). Less is more: Trading a little bandwidth for ultra-low latency in the data center. Em *Proceedings of USENIX NSDI conference*.

[Allan et al., 2010] Allan, D., Ashwood-Smith, P., Bragg, N., Farkas, J., Fedyk, D., Ouellete, M., Seaman, M. e Unbehagen, P. (2010). Shortest path bridging: Efficient control of larger Ethernet networks. *Communications Magazine, IEEE*, 48(10):128–135.

[Bari et al., 2013] Bari, M., Boutaba, R., Esteves, R., Granville, L., Podlesny, M., Rabbani, M., Zhang, Q. e Zhani, M. (2013). Data center network virtualization: A survey. *Communications Surveys Tutorials, IEEE*, 15(2):909–928.

[Benson et al., 2010] Benson, T., Akella, A. e Maltz, D. A. (2010). Network traffic characteristics of data centers in the wild. Em *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, IMC '10, p. 267–280. ACM.

[Costa et al., 2012] Costa, L. H., de Amorim, M. D., Campista, M. E. M., Rubinstein, M., Florissi, P. e Duarte, O. C. M. B. (2012). Grandes massas de dados na nuvem: Desafios e técnicas para inovação. Em *Minicursos do Simpósio Brasileiro de Redes de Computadores-SBRC 2012*.

[Couto et al., 2012] Couto, R. S., Campista, M. E. M. e Costa, L. H. M. K. (2012). A reliability analysis of datacenter topologies. Em *Global Telecommunications Conference (GLOBECOM 2012), IEEE*, p. 1 –6.

[Greenberg et al., 2011] Greenberg, A., Hamilton, J. R., Jain, N., Kandula, S., Kim, C., Lahiri, P., Maltz, D. A., Patel, P. e Sengupta, S. (2011). VL2: A scalable and flexible data center network. *Commun. ACM*, 54(3):95–104.

[Gude et al., 2008] Gude, N., Koponen, T., Pettit, J., Pfaff, B., Casado, M., McKeown, N. e Shenker, S. (2008). NOX: Towards an operating system for networks. Em *SIGCOMM Comput. Commun. Rev., 2008*, p. 105–110. ACM.

[Guimarães et al., 2013] Guimarães, P. H. V., Murillo P., A. F., Andreoni L., M. E., Mattos, D. M. F., Ferraz, L. H. G., Pinto, F. A. V., Costa, L. H. M. K. e Duarte, O. C. M. B. (2013). Comunicação em redes elétricas inteligentes: Eficiência, confiabilidade, segurança e escalabilidade. Em *Minicursos do Simpósio Brasileiro de Redes de Computadores - SBRC*, p. 101–164, Brasília, DF, Brazil.

[Mattos e Duarte, 2012] Mattos, D. M. F. e Duarte, O. C. M. B. (2012). QFlow: Um sistema com garantia de isolamento e oferta de qualidade de serviço para redes virtualizadas. Em *XXX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos - SBRC'2012*.

[McKeown et al., 2008] McKeown, N., Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, L., Rexford, J., Shenker, S. e Turner, J. (2008). OpenFlow: Enabling innovation in campus networks. *SIGCOMM Comput. Commun. Rev., 2008*.

[Mudigonda et al., 2010] Mudigonda, J., Yalagandula, P., Al-Fares, M. e Mogul, J. C. (2010). SPAIN: COTS data-center Ethernet for multipathing over arbitrary topologies. Em *Proceedings of the 7th USENIX NSDI conference*, NSDI'10. USENIX Association.

[ns3, 2006] ns3 (2006). The ns3 network simulator. `http://www.nsnam.org/`.

[Raiciu et al., 2011] Raiciu, C., Barre, S., Pluntke, C., Greenhalgh, A., Wischik, D. e Handley, M. (2011). Improving datacenter performance and robustness with multipath TCP. Em *Proceedings of the ACM SIGCOMM 2011 conference*, SIGCOMM '11, p. 266–277. ACM.

[Touch e Perlman, 2009] Touch, J. e Perlman, R. (2009). Transparent interconnection of lots of links (TRILL): Problem and applicability statement. RFC 5556 (Informational).

[Wilson et al., 2011] Wilson, C., Ballani, H., Karagiannis, T. e Rowtron, A. (2011). Better never than late: Meeting deadlines in datacenter networks. Em *Proceedings of the ACM SIGCOMM 2011 conference*, SIGCOMM '11, p. 50–61. ACM.

[Zats et al., 2012] Zats, D., Das, T., Mohan, P., Borthakur, D. e Katz, R. (2012). Detail: Reducing the flow completion time tail in datacenter networks. *SIGCOMM Comput. Commun. Rev.*, 42(4):139–150.