Compreendendo o Desempenho de Serviços Encadeados Virtuais de Redes

Alexandre Heideker, Ivan Zyrianoff, Carlos Kamienski

Universidade Federal do ABC (UFABC) – Santo André, SP – Brazil
{alexandre.heideker, ivan.dimitry, cak}@ufabc.edu.br

Abstract. *The concepts of Software Defined Networks (SDN) and Network Function Virtualization (NFV) have promoted network chaining, or Service Chaining(SC), quickly and simply. In dynamic infrastructure scenarios, the management of these virtual elements presents challenges both for the connection of these elements and for the evaluation of their states in elastic projects. Most of the works have dealt with this problem with probabilistic heuristics or AI-based techniques in static projects. This work presents the performance evaluation of SCs and their analytical understanding using Queueing Theory. The results obtained in this study allow the use of queuing theory not only as a validation technique for IaaS projects, but also as heuristic for dynamic evaluation in production environments.*

Resumo. *Os conceitos de Redes Definidas por Software (SDN) e Virtualização de Funções de Rede (NFV) promoveram o encadeamento de serviços de rede, ou Service Chaining (SC), de forma rápida e simples. Em cenários de infraestrutura dinâmica, a manipulação destes elementos virtuais apresenta desafios tanto para a conexão destes elementos como para a avaliação de seus estados em projetos elásticos. A maioria dos trabalhos tem atacado este problema com heurísticas probabilísticas ou baseadas em técnicas de Inteligência Artificial em projetos estáticos. Este trabalho apresenta a avaliação de desempenho de SCs e sua compreensão analítica utilizando Teoria das Filas. Os resultados obtidos neste estudo permitem utilizar a teoria das filas não só como técnica de validação para projetos de IaaS, como também heurística para avaliação dinâmica de ambientes de produção.*

1. Introdução

Ao longo dos últimos 10 anos, uma forte tendência de "softwarização" vem sendo observada na área de telecomunicações. Com a ampla adoção da computação em nuvem como paradigma de infraestrutura, diversas tecnologias estão sendo apresentadas no sentido de tornar eficiente e dinâmico o processo de migração destes recursos para a nuvem, com especial destaque às Redes Definidas por Software (SDN) [McKeown et al. 2008] e a Virtualização de Funções de Rede (NFV) [ETSI et al. 2012]. O encadeamento destas Funções de Rede (NF) através de *middleboxes* virtualizados e tecnologias de enlace clássicas ou definidas por software dão origem ao conceito de Serviço Encadeado ou *Service Chaining* (SC).

Entre as vantagens preconizadas por estas tecnologias, a manipulação destes recursos via software, de forma dinâmica e automatizada, apresentam um novo

horizonte de desafios para pesquisadores e indústria, em especial na orquestração otimizada destes recursos, com reflexos diretos na qualidade de serviço e custo de operação, este último próprio do modelo de negócios da computação em nuvem, ou seja, pague pelo que usar pelo tempo que usar. No centro deste desafio, a criação de modelos analíticos, estocásticos e híbridos faz-se necessário para desenvolver softwares reativos ou preditivos, permitindo que requisitos cada vez mais exigentes sejam corretamente atendidos, em projetos estáticos ou elásticos.

A maioria dos trabalhos tem abordado este problema considerando cenários estáticos com requisitos bem definidos, atendendo as mais diversas restrições e utilizando técnicas de Inteligência Artificial (IA) e Aprendizado de Máquina. Apesar de obterem resultados satisfatórios, há restrições para o uso deste tipo de técnica, seja por questões relacionadas ao tempo de resposta e assertividade da técnica utilizada, ou pela dificuldade em validar modelos de comportamento fechados, conhecidos como "caixa preta".

Este trabalho apresenta a avaliação experimental de SCs e sua compreensão analítica utilizando Teoria das Filas, uma técnica clássica para modelar centros de serviço, permitindo a avaliação imediata do estado do sistema, projetar e gerenciar de forma autônoma SCs, bem como a escolha de diferentes variações para otimização de desempenho e custo. Os resultados obtidos nos experimentos mostram comportamentos similares entre os resultados teóricos e empíricos, demonstrando a assertividade da técnica. Além disso, o estudo não utiliza o uso de CPU como métrica de avaliação, abordagem clássica deste tipo de estudo, permitindo que a técnica seja aplicada em cenários em que o uso de CPU não possui correlação direta com o estado das funções de rede.

A seção 2 apresenta os conceitos relacionados aos SC virtualizados e tecnologias correlatas. A seção 3 apresenta os trabalhos correlatos. A seção 4 traz a metodologia utilizada e os experimentos preliminares para definição de parâmetros e métricas. Na seção 5 são apresentados os resultados obtidos pelos experimentos assim como as discussões sobre os resultados obtidos e lições aprendidas.

2. Conceitos Básicos

O modelo de referência OSI, assim como o próprio modelo TCP/IP, têm como objetivo a criação de camadas de abstração da infraestrutura de rede necessária à comunicação em rede de computadores [Sousa et al. 2007]. Essa abstração mostrou-se eficaz e promoveu o crescimento das redes de computadores. No cenário atual, é comum que a informação, entre sua origem e seu destino, atravesse um grande número de elementos de rede responsáveis por este trajeto. Além dos enlaces, estão envolvidos nesta tarefa os chamados *middleboxes*, entre os quais destacam-se Roteadores, *Network Address Translator* (NAT), *Virtual Private Network* (VPN), *Proxy*, *Firewall*, *Load Balancer* (LB), entre outros.

Segundo Patouni [Patouni et al. 2013], a virtualização de redes consiste na abstração destes recursos, simplificando a tarefa de alocação, considerando localização física e uso, além de promover a separação entre recursos físicos e lógicos.

As Redes Definidas por Software(SDN)[McKeown et al. 2008] representam uma abordagem completamente inovadora na área de redes de computadores. A

proposta é separar os planos de controle e de dados, centralizando a inteligência e o estado da rede, abstraindo a infraestrutura de rede das aplicações. Sob esse ponto de vista, o hardware de rede passa a ser considerado apenas o meio físico no qual a informação flui, transferindo as decisões sobre esses fluxos de dados para a figura do controlador. De forma mais abrangente, essa tecnologia permite a interferência direta do desenvolvedor no fluxo da rede, criando assim a oportunidade de criar aplicações que manipulam a rede física da mesma forma que uma rede virtual.

O trajeto percorrido pela informação nas redes atuais, como foi mencionado, envolve outros elementos além dos enlaces. Estes *middleboxes*, sob o ponto de vista da virtualização de redes, são denominados Funções de Rede (NF). Essas funções de rede são, via de regra, implementadas por equipamentos dedicados, com hardware e software proprietários, conhecidos como *appliances*. De acordo com o ETSI, além do alto custo de aquisição destes equipamentos, as tecnologias utilizadas impedem a evolução e/ou modificação destes para implementar novas ideias e tecnologias experimentais. Do ponto de vista do gerenciamento, a tarefa é por vezes realizada apenas no local ou via console, além de exigir modificações nas conexões físicas em certas circunstâncias. Finalmente, o dimensionamento destes é realizado considerando a demanda máxima, gerando desperdícios na aquisição dos equipamentos e no consumo de energia. Quando esta demanda máxima é superada, novos equipamentos devem ser adquiridos, configurados e instalados no local, gerando um tempo de solução para o problema da demanda que pode compreender algo entre horas ou até dias.

Considerando este cenário e com as novas exigências e paradigmas como a Computação em Nuvem, o *European Telecommunications Standards Institute* (ETSI) propôs em 2012 o conceito de *Network Function Virtualization* (NFV), consistindo na substituição de equipamentos dedicados por versões virtuais suportadas por servidores e equipamentos de rede commodities utilizando técnicas de virtualização.

Apesar de não haver uma relação de dependência entre NFV, SDN e Computação em Nuvem, em [Mijumbi et al. 2015; Open Networking Foundation 2015], há uma relação de benefício mútuo na intersecção entre NFV e SDN, promovendo a automação, isolação e agilidade entre NFV e Computação em Nuvem na orquestração, elasticidade e provimento de recursos. A intersecção das três tecnologias permite a implementação do conceito de Infraestrutura como Serviço (IaaS) de forma ampla e eficiente.

Como já mencionado, o trajeto nas redes atuais envolve um ou mais elementos em seu trajeto e o conjunto destes enlaces e NFs é denominado *Service Chaining* (SC). A utilização de tecnologias de virtualização e definição por software nestes dispositivos nos permite a construção, gerenciamento e manipulação automatizada deste conjunto de elementos. Neste contexto, um desafio observado em [Khalid et al. 2016] é a necessidade de definir não só os elementos que compõem o SC, mas também a ordem específica na qual a informação deve fluir, como um grafo direcionado. As heurísticas utilizadas para otimização destes SCs com múltiplas restrições podem ser beneficiadas com o auxilio de técnicas analíticas para redução na complexidade do problema.

A Teoria das Filas é uma técnica clássica para analise de centros de serviço, na qual o sistema pode ser modelado como um ou mais servidores (neste contexto sem

uma relação direta com o hardware) e uma ou mais filas com tarefas que aguardam para serem executadas [Jain et al. 1991]. Essa teoria pode ser facilmente aplicada à cenários computacionais e de telecomunicações, nos quais servidores podem ser considerados não só como o hardware de um servidor como uma CPU ou um *middlebox*. As filas nada mais são que os diversos buffers encontrados em inúmeros dispositivos digitais.

Além do servidor e da fila, que são caracterizados em seu tamanho, quantidade e distribuição, outros dois elementos são essenciais para este tipo de análise: a distribuição de chegada e a distribuição de serviço. A distribuição de chegada define com as novas tarefas, mensagens, pacotes ou fluxos de dados chegam ao sistema. Já a distribuição de serviço define o tempo necessário para completar cada tarefa que chega ao sistema, como o tempo necessário para encaminhar um pacote, uma mensagem ou processar uma tarefa.

3. Trabalhos Relacionados

Desde o início da adoção do paradigma de computação em nuvem, o processo de elasticidade de serviços, ou auto escala tem sido alvo de diversos estudos, entre eles o trabalho de Suleiman [Suleiman and Venugopal 2013] apresenta o processo de elasticidade de serviços de nuvem com base na teoria das filas, utilizando a métrica do use de CPU para reagir à demanda. Este trabalho diferencia-se por não considerar o uso de CPU com uma métrica geral, considerando que diversas funções de rede não afetam diretamente a CPU quando estão sobrecarregados, seja pela característica do serviço, seja pela própria largura de banda disponível para o fluxo de dados.

Em [Mijumbi et al. 2015], o desafio da orquestração de Funções de Rede Virtualizadas (VNF) é abordado examinando o estado da arte dos ambientes disponíveis para implementação da tecnologia NFV. O trabalho apresenta como grandes desafios para a área, entre outros, a estratégia de alocação de recursos e a programabilidade e interação com estes SCs, desafio que pode ser muito beneficiado com o arcabouço analítico apresentado neste trabalho.

Considerando as aplicações envolvendo telefonia celular de quinta geração (5G), em [Blanco et al. 2017] a virtualização de redes é colocado com um dos pilares da tecnologia que apoia o 5G, e que desafios como a melhor distribuição das VNFs na infraestrutura disponível, as possíveis interferências entre estes diversos dispositivos virtualizados e atender à requisitos funcionais simultaneamente, pode ser considerado um problema Não Polinomial Difícil (NP-hard). A possibilidade de soluções analíticas para algumas dessas demandas, explorada neste trabalho, podem reduzir expressivamente a complexidade destes problemas.

Trabalhos como [Bremler-barr [S.d.]; Császár et al. 2013; Ge et al. 2014] apresentam diversas plataformas e frameworks para fornecimento e gerenciamento de VNFs e SCs porém sem considerar aspectos dinâmicos durante o uso destes elementos virtualizados, processos de sensoriamento e elasticidade destes SCs.

Em [Khalid et al. 2016], a proposta de uma interface padrão de sensoriamento de VNFs, mapeando métricas e criando uma API padronizada mostra-se fundamental para a ampla adoção da tecnologia NFV, permitindo que algoritmos possam obter informações em tempo real do estado destas VNFs e que heurísticas sejam aplicadas

para o correto gerenciamento da infraestrutura virtualizada – neste ponto, abordagens analíticas podem contribuir fortemente neste cenário. Também considerando aspectos de sensoriamento de VNFs e a aplicação de NFV em um caso de uso real, em [Heideker and Kamienski 2016] os métodos de sensoriamento são discutidos e um algoritmo de elasticidade é implementado de forma reativa apenas considerando o estado atual das VNFs – neste caso, o uso de técnicas analíticas permite não só otimizar o mecanismo de elasticidade, como também implementar características preditivas neste algoritmo.

4. Metodologia

A metodologia deste trabalho está dividida em três etapas: experimentos preliminares para detecção das características das diferentes VNFs e do SC avaliado; avaliação experimental de duas possíveis variações na configuração deste SC; e finalmente a aplicação da Teoria das Filas para compreensão analítica dos resultados. O cenário utilizado considera o típico SC em uma rede de dados, exibido na Figura 1, com um Firewall e um NAT entre a Internet.

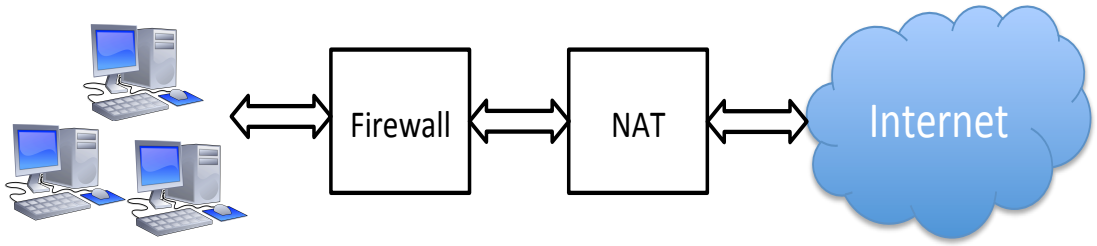

Figura 1: Cenário Utilizado para o Modelo.

Todos os cenários experimentais utilizam duas ferramentas para avaliação de desempenho: o iPerf para obter o desempenho máximo dos cenários e um gerador de tráfego estocástico desenvolvido especificamente para obter as métricas de vazão e latência observadas pelo usuário. Para detectar com precisão a latência e a vazão durante os experimentos, este gerador produz um tráfego com requisições exponencialmente espaçadas e com tamanhos seguindo uma distribuição Lognormal com tamanho médio de 5Kbytes (*Payload*). Cada requisição do gerador de tráfego é obtida pela abertura de um *socket* TCP entre o Gerador de Tráfego e o Alvo e a transferência do *Payload* entre eles através deste *socket*. A Figura 2 apresenta a estrutura de dados utilizada pelo gerador de tráfego para obter as métricas entre o Gerador de Tráfego (A) e o Alvo (B). O Gerador de Tráfego preenche parte desta estrutura (id, Saída de A, Tamanho, *Payload*, Chegada em A e Fim da transferência), que é completada pelo Alvo (Chegada em B, Fim da Transferência), que devolve o pacote ao gerador de tráfego para posterior cálculo de latência, vazão e tempo de serviço.

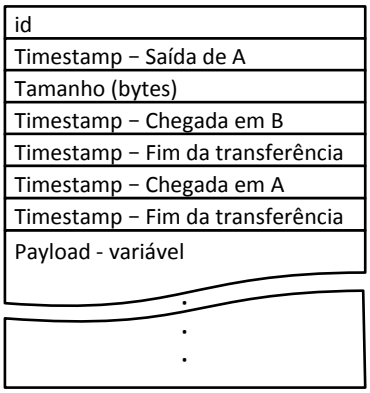

Figura 2: Estrutura do Bloco de Dados Enviado ao Alvo.

A implementação do NAT utilizado nos experimentos utiliza o módulo Iptables do Linux, assim como o Firewall, que conta com 100mil regras de rejeição de endereços IP em sua configuração. Os experimentos foram realizados em uma máquina física utilizando o processador Intel Core i7 @ 3,1 GHz com 16Gbytes de memória RAM DDR3 e Linux Ubuntu 16.04 LTS 64bits. A configuração dos diferentes cenários foram realizadas utilizando sub-redes configuradas em um switch virtual Open vSwitch 2.5.2.

Os experimentos foram realizados utilizando containers LXC para obter dados precisos de latência e taxa de serviço, aproximando os resultados de abordagens simuladas. Como em um ambiente experimental baseado em containers há o compartilhamento direto do mesmo *kernel* do sistema operacional hospedeiro, o mesmo relógio também é compartilhado, evitando desta forma a necessidade de sincronização e evitando a influência do erro presente em protocolos com o NTP.

A configuração experimental da primeira etapa de avaliações, exibida na Figura 3, visa obter as métricas de latência, tempo de serviço e vazão do NAT, do Firewall (FW) e do SC completo (FW+NAT) em diferentes condições de taxa de requisição (taxas variando de 100 à 1200 requisições por segundo em incrementos de 100).

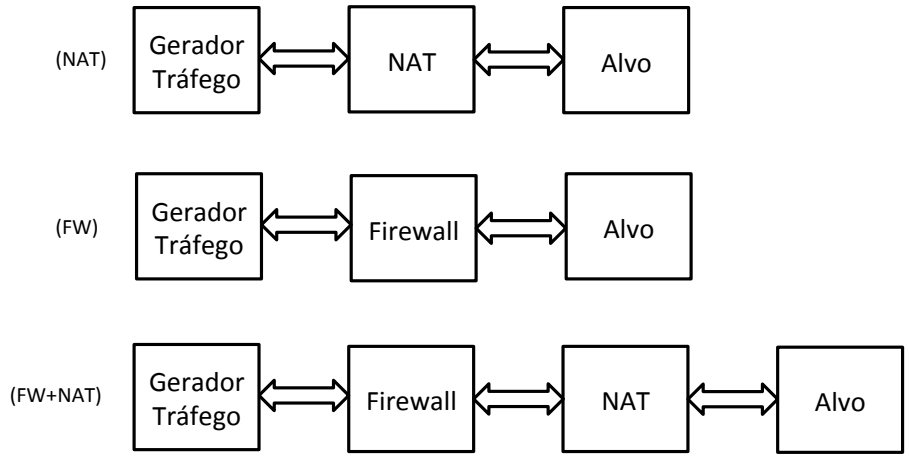

Figura 3: Esquemas Utilizados para Avaliação Individual das Funções de Rede e do Serviço Encadeado.

Com o desempenho obtido nestes experimentos preliminares, um novo conjunto de experimentos utiliza o conhecimento obtido para a avaliação do cenário completo, além da compreensão de duas possíveis técnicas para ampliar a capacidade do SC. A Figura 4 mostra duas possíveis modificações no cenário FW+NAT da Figura 2: 2FW+2NAT no qual o SC é duplicado e a configuração 2FW+NAT, que duplica apenas a função de rede que está sobrecarregada, ou seja, o firewall.

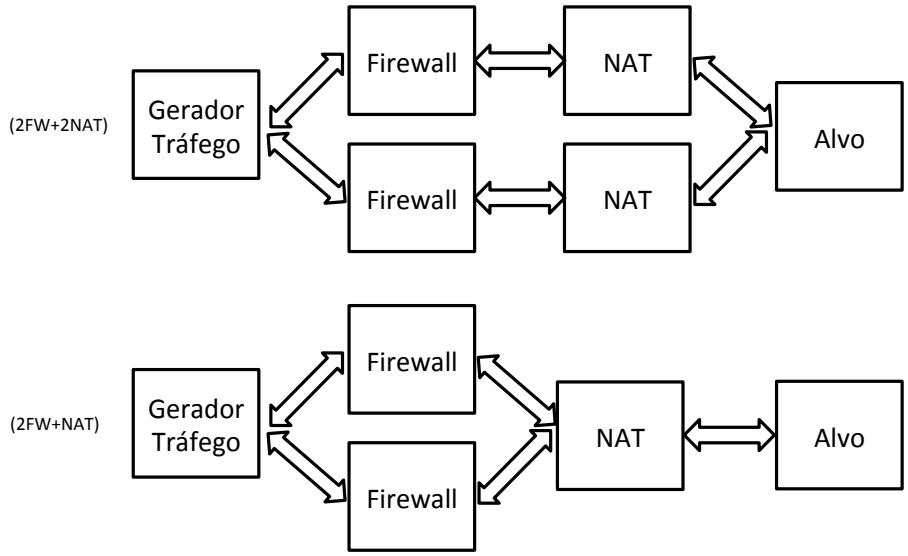

Figura 4: Variações do Cenário FW+NAT para Ampliar a Capacidade do SC.

Finalmente, através dos dados obtidos nos experimentos, é possível utilizar a Equação (1) de intensidade de tráfego para filas do tipo M/M/m para identificar a viabilidade do sistema, a qual deve ser menor ou igual a 1.

$$\boldsymbol{\rho = \frac{\lambda}{m\mu}} \tag{1}$$

onde:

$\lambda \rightarrow taxa\ de\ requisição$

$m \rightarrow número\ de\ servidores$

$\mu \rightarrow taxa\ de\ serviço$

Outra métrica importante obtida através da Teoria da Filas é a probabilidade de não haver fluxos no sistema, ou seja, o sistema estar ocioso através da Equação (2). Em condições de congestionamento de trafego, o número de fluxos no sistema tende a subir, ou seja, há uma probabilidade de enfileiramento no sistema, probabilidade esta obtida pela Equação (3). Já a Equação (4) apresenta probabilidade de um dado número de fluxos estarem no sistema simultaneamente.

$$\boldsymbol{p_0 = \left[1 + \frac{(m\rho)^m}{m!\,(1-\rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!}\right]^{-1}} \tag{2}$$

$$\boldsymbol{\varrho = \frac{(m\rho)^m}{m!\,(1-\rho)}\, p_0} \tag{3}$$

$$p_n = \begin{cases} p_0 \frac{(m\rho)^n}{n!}, & n < M \\ p_0 \frac{\rho^n m^m}{m!}, & n \geq M \end{cases} \quad (4)$$

onde:

$$n \rightarrow número\ de\ fluxos\ no\ sistema$$

E finalmente, com os resultados obtidos nas Equações (2) e (3), é possível calcular o tempo médio de resposta do sistema, através da Equação (5).

$$E = \frac{1}{\mu}\left(1 + \frac{\varrho}{m(1-\rho)}\right) \quad (5)$$

5. Resultados e Discussões

5.1. Experimentos Preliminares

Os cenários NAT, FW e FW+NAT foram submetidos a taxas de requisições médias de 100 à 1.200 requisições por segundo, com incremento de 100 requisições por segundo. Os resultados obtidos podem ser vistos na Figura 5. É possível observar que os cenários FW e FW+NAT possuem um comportamento estável até serem submetidos à uma taxa de 700 requisições por segundo. Após este patamar, o mecanismo de controle de congestionamento do protocolo TCP começa a modificar o comportamento do sistema demonstrando que o tráfego está próximo da capacidade máxima de processamento de pacotes das VNFs.

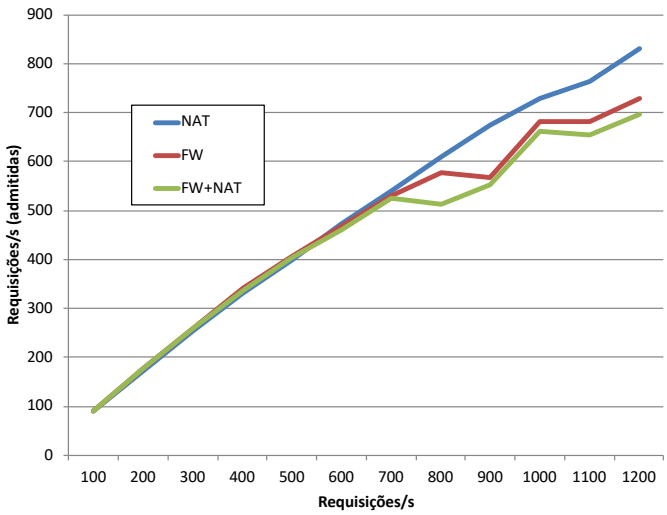

Figura 5: Taxa de requisições admitidas em relação à taxa de requisições submetidas ao sistema nos cenários NAT, FW e FW+NAT.

Considerando que os resultados obtidos são estáveis até 700 requisições por segundo, a Tabela 1 mostra os valores obtidos para a taxa média de serviço, com intervalos de confiança de 95%, obtido após 15 repetições. Utilizando a Equação (1) e a taxa de serviço obtida experimentalmente, a intensidade do tráfego é vista na Figura 6 demonstrando a viabilidade do sistema garantida pelo valor de $\rho<1$.

Tabela 1: Tempo médio de serviço e taxa média de serviço

	Tempo de Serviço (milissegundos)	Taxa de Serviço – μ (fluxos por segundo)
NAT	0,264 ± 0,012	3806 ± 166
FW	1,840 ± 0,074	536 ± 22
FW+NAT	1,998 ± 0,066	484 ± 16

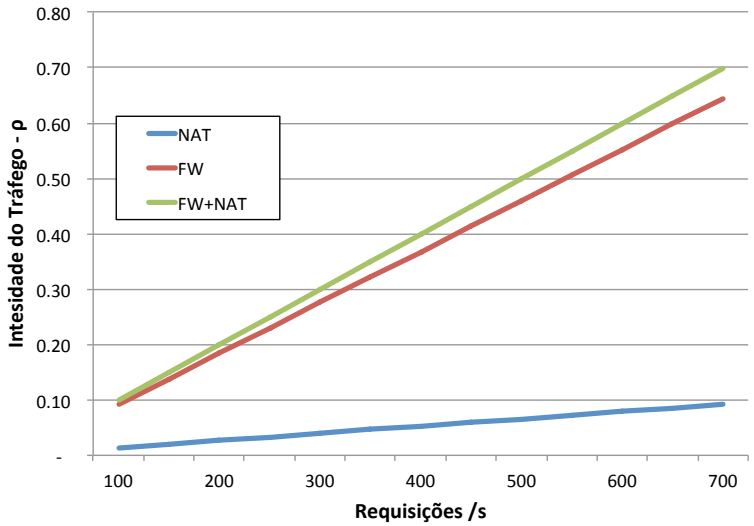

Figura 6: Intensidade do tráfego teórica obtidas pela Equação (1).

Os valores obtidos pelas métricas de vazão e latência considerando a faixa de requisições entre 100 e 700 requisições por segundo são exibidas na Tabela 2, com intervalos de confiança de 95%, obtidos após 15 repetições. A diferença dos valores de vazão observada entre a obtida pelo tráfego exponencial e o obtido pelo iPerf , deve-se ao fato deste último ter por objetivo obter a máxima vazão entre dois pontos e não considerar um comportamento real de tráfego, ou seja, requisições exponencialmente espaçadas com tamanhos seguindo uma distribuição Lognormal. Além disso, o valor de vazão observado nos experimentos é a média de cada transferência efetuada, representando desta forma percepção de um usuário individual neste sistema.

Tabela 2: Vazão e Latência obtida nos experimentos.

	Vazão (Exponencial) Mbytes/s	Vazão (iPerf) Mbytes/s	Latência microssegundos
NAT	150,3 ±8,7	563,9 ±3,9	155 ±4,8
FW	17,5 ±0,45	25,4 ±1,6	753 ±76,7
FW+NAT	16,0 ±0,28	27,5 ±2,1	913 ±83,0

Utilizando agora a Equação (4), obtemos o tempo de resposta teórico do sistema, ou seja, o tempo de serviço esperado. A Figura 7 mostra este valor teórico juntamente com o valor obtido experimentalmente. É possível observar uma grande similaridade entre o valor teórico e experimental no cenário NAT e uma diferença

muito pequena entre os valores teóricos e experimentais no caso do cenário FW e FW+NAT.

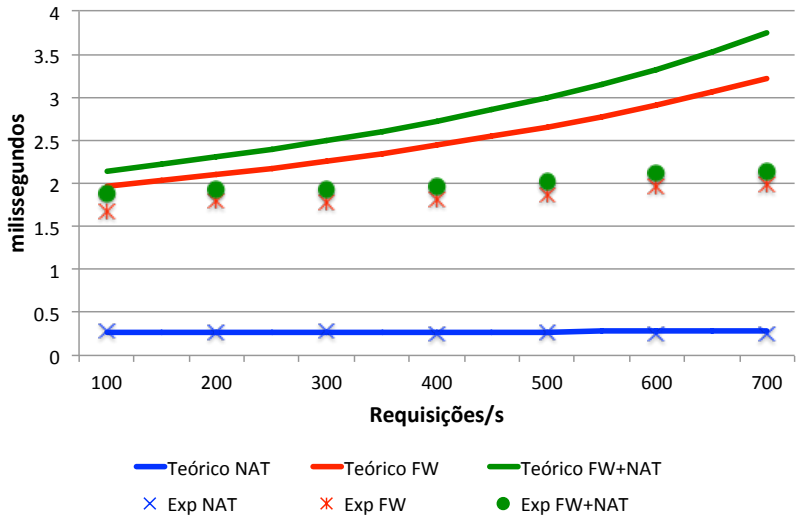

Figura 7: Tempo de resposta teórico do sistema comparado com o valor obtido experimentalmente nos cenários NAT, FW e FW+NAT.

5.2. Avaliação do Serviço Encadeado

Com os resultados obtidos nos experimentos preliminares, tem-se a avalição de três diferentes configurações de um SC, ou seja, os cenários FW+NAT, 2FW+2NAT e 2FW+NAT. O primeiro passo é obter experimentalmente a capacidade do SC em admitir conexões antes que o mecanismo de controle de congestionamento do protocolo TCP atue. A Figura 8 mostra um comportamento similar ao observado nas funções de rede individuais, obtendo-se mais uma vez um limite próximo de 700 requisições por segundo após o qual o sistema torna-se instável.

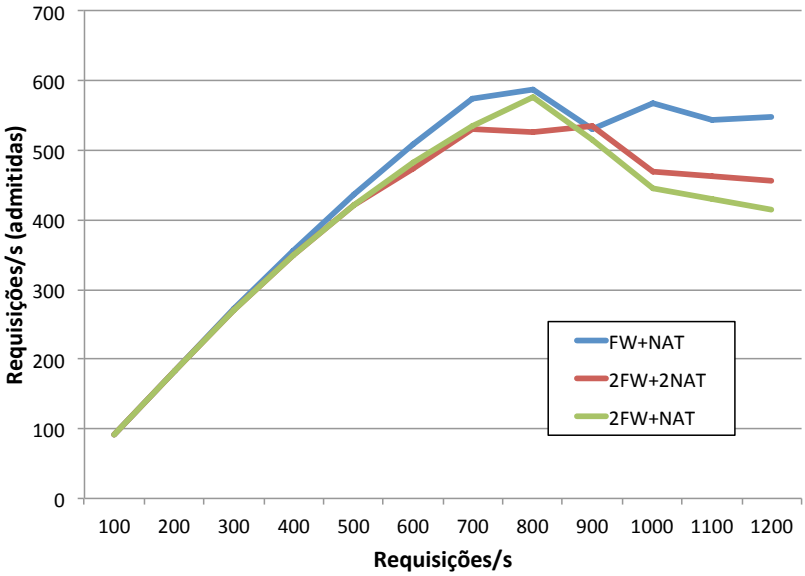

Figura 8: Taxa de requisições admitidas em relação à taxa de requisições submetidas ao sistema nos cenários FW+NAT, 2FW+2NAT e 2FW+NAT.

O mesmo comportamento anômalo pode ser observado na métrica de vazão do sistema, exibida na Figura 9(a), com a queda acentuada do desempenho após 700 requisições por segundo. Na Figura 9(b) também pode-se observar o mesmo comportamento, agora sob a métrica de latência.

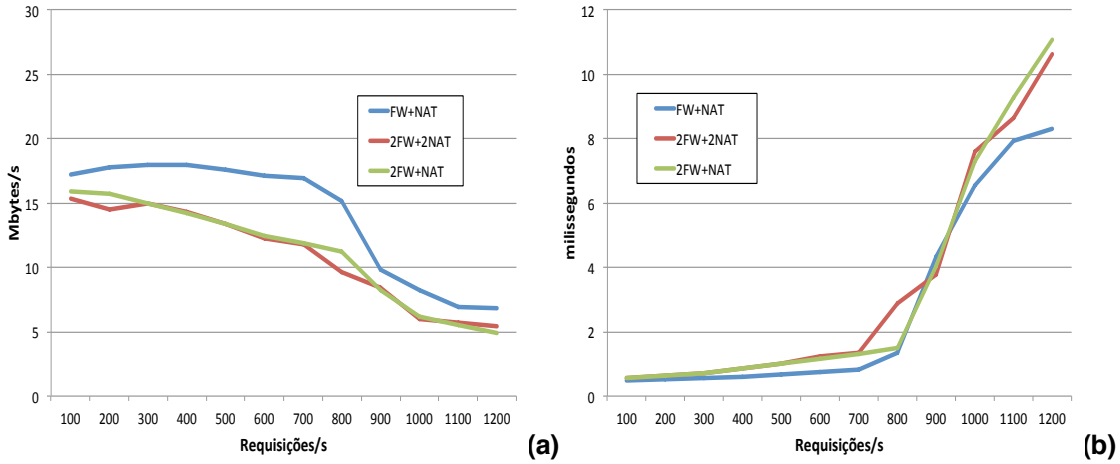

Figura 9: (a)Vazão total obtida nos experimentos FW+NAT, 2FW+2NAT e 2FW+NAT e (b)Latência obtida nos mesmos experimentos.

Aplicando a Equação (1), obtemos a curva de intensidade do tráfego, exibida na Figura 10. Segundo a Teoria das Filas, o sistema é viável apenas se a intensidade do tráfego obtida for menor que 1, e pode-se observar que, nos três cenários, a intensidade do tráfego começa a aproximar-se deste valor após ultrapassar a faixa de 700 requisições por segundo, mostrando mais uma vez a presença deste limite. Como após aproximar-se deste valor o mecanismo de controle de congestionamento começa atuar sobre os protocolos de rede, o aumento do atraso no estabelecimento de uma conexão impede a admissão de novas conexões, reduzindo dessa forma que o sistema entre em colapso. Apesar dessa característica intrínseca do protocolo TCP, a qualidade de serviço observada pelo usuário é comprometida, observada pela redução na vazão e aumento da latência.

Considerando apenas o intervalo de 100 até 700 requisições por segundo, a Figura 11(a) mostra o resultado obtido com a Equação (2) para a probabilidade de ociosidade do sistema e na Figura 11(b) tem-se o resultado obtido com a Equação (3) para a probabilidade de enfileiramento de fluxos no sistema. A escolha deste intervalo de requisições baseia-se no fato que o sistema só apresenta viabilidade teórica quando a intensidade do tráfego é menor que 1.

Finalmente, utilizando a Equação (5) obtemos o tempo de resposta do sistema. A Figura 12 mostra os resultados obtidos com o uso da equação sobre os resultados obtidos experimentalmente, mostrando um comportamento similar entre os resultados teóricos e experimentais nos três cenários.

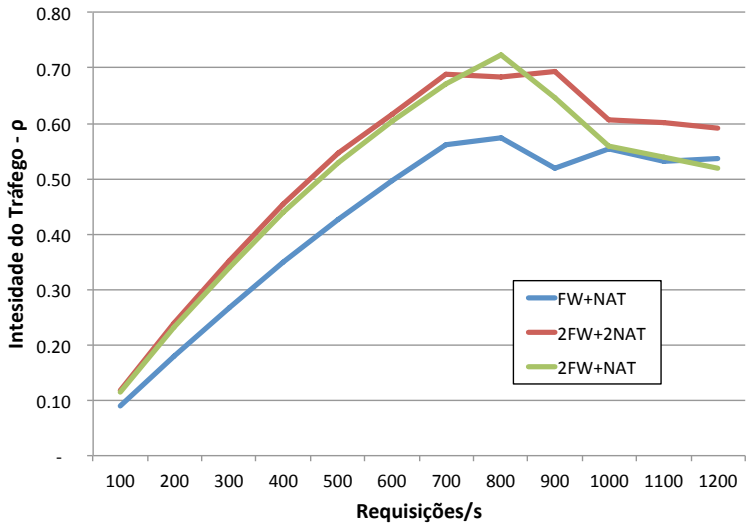

Figura 10: Intensidade do tráfego nos experimentos FW+NAT, 2FW+2NAT e 2FW+NAT.

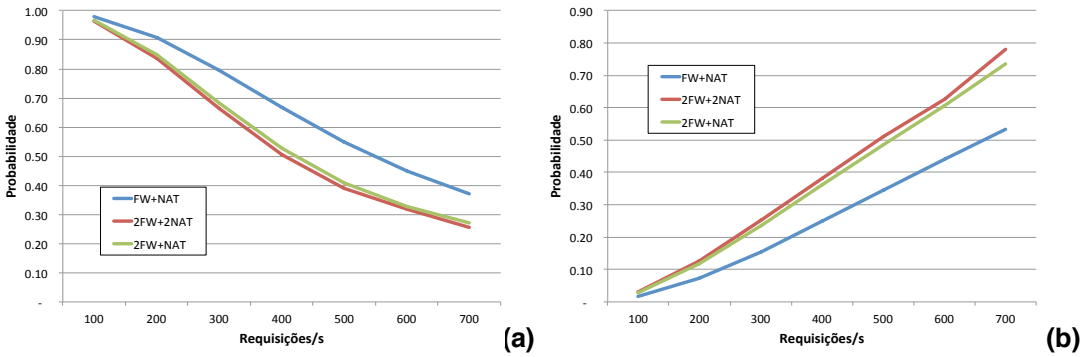

Figura 11: (a)Distribuição de probabilidade de ociosidade no sistema em relação à taxa de requisições nos cenários FW+NAT, 2FW+2NAT e 2FW+NAT e (b)Distribuição de probabilidade de enfileiramento de fluxos no sistema em relação à taxa de requisições nos cenários FW+NAT, 2FW+2NAT e 2FW+NAT.

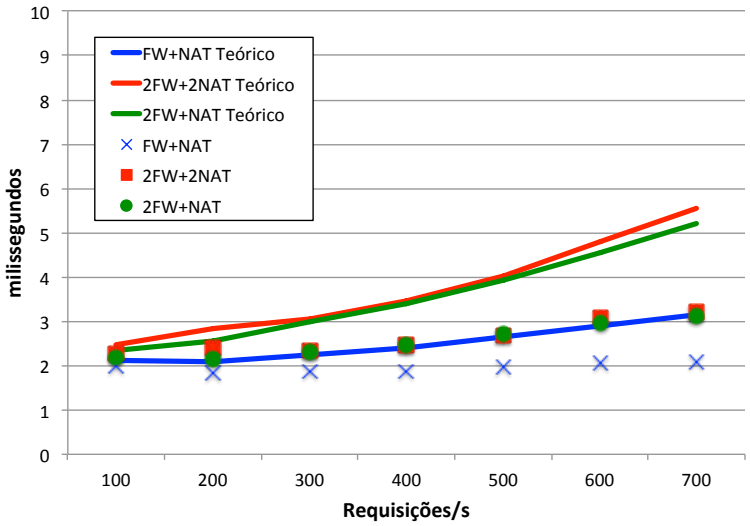

Figura 12: Tempo de resposta teórico do sistema comparado com o valor obtido experimentalmente nos cenários FW+NAT, 2FW+2NAT e 2FW+NAT.

Apesar do objetivo de representar experimentalmente o ambiente real, a implementação do experimento com o uso de máquinas virtuais em uma mesma máquina real não produziu os resultados esperados, já que é intuitivo que o cenário 2FW+2NAT e 2FW+NAT deveriam apresentar um desempenho bruto melhor que o cenário FW+NAT, já que a tarefa de filtrar, modificar e encaminhar os fluxos de dados é paralelizada. Este comportamento discrepante pode ser explicado por fatores relacionados ao compartilhamento de estruturas de dados no *kernel* do sistema operacional durante o uso de containers, comportamento este também observado em [Heideker and Kamienski 2016]. Além disso, a abordagem baseada em containers para implementar VNFs tem despertado o interesse, principalmente em ambientes com pouco poder computacional, como é o caso de pequenos dispositivos na borda da rede, como em [Cziva et al. 2016] e [Cziva and Pezaros 2017]. Este fato não compromete a análise do SC através da teoria das filas já que o comportamento geral do sistema, em todos os cenários avaliados, apresenta uma boa correlação com os resultados teóricos.

6. Conclusões

Os resultados obtidos nos experimentos permitiram identificar claramente os elementos necessários ao uso da teoria das filas para analise de um *service chaining* virtual, permitindo não só prever o desempenho do sistema, como também identificar possíveis gargalos e, onde a aplicação necessite de um mecanismo de elasticidade, atuar para garantir sempre a melhor relação custo/benefício para a infraestrutura.

Outra importante observação é a equivalência teórica e, como foi comprovado nos experimentos, prática, do processo de elasticidade do SC ser por função sobrecarregada ou por completo, avaliado nos cenários 2FW+2NAT e 2FW+NAT. Esta conclusão permite que a redução no número de instancias utilizadas no SC gere uma economia na operação.

Como trabalhos futuros, espera-se implementar este experimento em escalas maiores minimizando desta forma a influência do erro de sincronização, além de submeter à teoria diferentes tipo de funções de rede.

7. Agradecimentos

Essa pesquisa foi parcialmente financiada pelo projeto SWAMP [Kamienski et. al. 2018], uma colaboração entre Brasil e União Europeia.

Referencias

Blanco, B., Fajardo, J. O., Giannoulakis, I., et al. (2017). Technology pillars in the architecture of future 5G mobile networks: NFV, {MEC} and {SDN}. Computer Standards & Interfaces, v. 54, n. December 2016.

Bremler-barr, A. ([S.d.]). OpenBox : A Software-Defined Framework for Developing , Deploying , and Managing Network Functions. p. 511–524.

Császár, A., John, W., Kind, M., et al. (2013). Unifying cloud and carrier network: EU FP7 Project UNIFY. Proceedings - 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing, UCC 2013, p. 452–457.

Cziva, R., Jouet, S., White, K. J. S. and Pezaros, D. P. (2016). Container-based network function virtualization for software-defined networks. In Proceedings - IEEE Symposium on Computers and Communications.

Cziva, R. and Pezaros, D. P. (2017). Container Network Functions: Bringing NFV to the Network Edge. IEEE Communications Magazine, v. 55, n. 6, p. 24–31.

ETSI, Chiosi, M., Clarke, D., et al. (2012). Network Functions Virtualisation. Citeseer, n. 1, p. 1–16.

Ge, X., Liu, Y., Du, D. H. C., et al. (2014). OpenANFV. Proceedings of the 2014 ACM conference on SIGCOMM - SIGCOMM '14, p. 353–354.

Heideker, A. and Kamienski, C. (2016). Gerenciamento Flexível de Infraestrutura de Acesso. XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos,

Jain, R., Simulation, M., Publishing, W. C. and Wiley, J. (1991). The art of computer systems performance analysis.

Kamienski, C., Soininen, J.P., Taumberger, M., Fernandes, S., Toscano, A., Salmon Cinotti, T., Filev Maia, R. Torre Neto, A., "SWAMP: an IoT-based Smart Water Management Platform for Precision Irrigation in Agriculture", aceito para o Global IoT Summit 2018 (GIoTS'18), Junho de 2018.

Khalid, J., Coatsworth, M., Gember-Jacobson, A. and Akella, A. (2016). A Standardized Southbound API for VNF Management. In Proceedings of the 2016 Workshop on Hot Topics in Middleboxes and Network Function Virtualization. . http://doi.acm.org/2940147.2940156.

McKeown, N., Anderson, T., Balakrishnan, H., et al. (2008). OpenFlow. ACM SIGCOMM Computer Communication Review, v. 38, n. 2, p. 69.

Mijumbi, R., Serrat, J., Gorricho, J., et al. (2015). Management and Orchestration Challenges in Network Function Virtualization. n. 3, p. 1–8.

Mijumbi, R., Serrat, J., Gorricho, J. L., et al. (2015). Network Function Virtualization: State-of-the-art and Research Challenges. n. c, p. 1–28.

Open Networking Foundation (2015). Relationship of SDN and NFV. . https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/onf2015.310_Architectural_comparison.08-2.pdf.

Patouni, E., Merentitis, A., Panagiotopoulos, P., Glentis, A. and Alonistioti, N. (2013). Network Virtualisation Trends: Virtually Anything Is Possible by Connecting the Unconnected. 2013 IEEE SDN for Future Networks and Services (SDN4FNS), p. 1–7.

Sousa, J. R. B., Sausen, P. S., Lima, a. M. N. and Perkusich, a. (2007). Redes de petri híbridas diferenciais: aplicação na modelagem e no gerenciamento dinâmico de energia de redes de sensores sem fio. Sba: Controle & Automação Sociedade Brasileira de Automatica, v. 18, n. 3, p. 278–291.

Suleiman, B. and Venugopal, S. (2013). Modeling performance of elasticity rules for cloud-based applications. Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOC, p. 201–206.