International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 12 | Dec 2018 www.irjet.net p-ISSN: 2395-0072

A Literature Survey on Scaling Approaches for VNF in NFV Monitoring

Shalmon Waidande

Pune, Maharashtra, India

Abstract - *The future of the Mobile Networks is based on the Virtual Network Functions (VNF). The new technologies Network Function Virtualization (NFV) and Software Defined Network (SDN) are enabling telecommunication operators for quick provisioning of the network services using VNFs. Although deployment and configuration of these VNFs over the cloud infrastructure is been maturing with cloud orchestration solutions, monitoring of these VNFs over their lifespan is still a challenge. Scaling of VNFs at right time is one of the challenge in monitoring scenarios to provide service assurance. This paper is literature survey of available architectures of monitoring the VNFs deployed on cloud infrastructure. Paper provides the comparative study of scaling mechanisms used in monitoring of VNFs and suggest a novel adaptive machine learning based approach for future work.*

Key Words: Network Function Virtualization, Monitoring, Virtual Machines, Auto-scaling

1. INTRODUCTION

Bringing in new services into traditional networks was becoming increasingly tough due to the proprietary nature of existing hardware devices, the cost of providing the space and energy for a range of middle boxes, and the shortage of skilled technicians to configure and maintain such services. With the success of the cloud computing architecture, where computing services are moved from end user devices to hypervisors located in Data-Center, the networking domain started increasing demand to move the telecom operator's infrastructure of network functions to such cloud platforms [3].

Network Functions Virtualization was then introduced to reduce these problems, along with other new technologies, such as Software Defined Networking (SDN) and Cloud Computing. NFV enables service providers and operators to abstract network service, like firewall, router and load balancer into software that functions on generic server. Currently, most vendors have started providing virtualised appliances and dumb boxes as Virtual Network Function (VNF) based service.

There are few challenges in NFV related to performance, reliability and security of the VNFs out of which scaling related challenges are discussed in next section of paper. Although the application of NFV as network services, benefits providers with automation, scalability, orchestration and cost-saving capacities, it brings several critical challenges while dealing with large-scale deployment and management of VNFs on a cloud infrastructure [5].

Next sections in papers describes available monitoring frameworks and gives comparative study of auto-scaling mechanisms being used in VNF monitoring.

1.1 VIM Monitoring

According to the architectural Framework for NFV [1], as established by ETSI, the Virtualized Infrastructure Management (VIM) is an essential component of the MANO (Management and Orchestration) platform, responsible for managing and monitoring the NFV Infrastructure (NFVI), including the physical and virtual entities. The main objective of VIM-level monitoring is to collect and aggregate metrics and events from physical and virtual resources and communicate these metrics in turn to the Orchestrator and the OSS/BSS. Fortunately, there are several technological solutions available which can be used towards this aim.

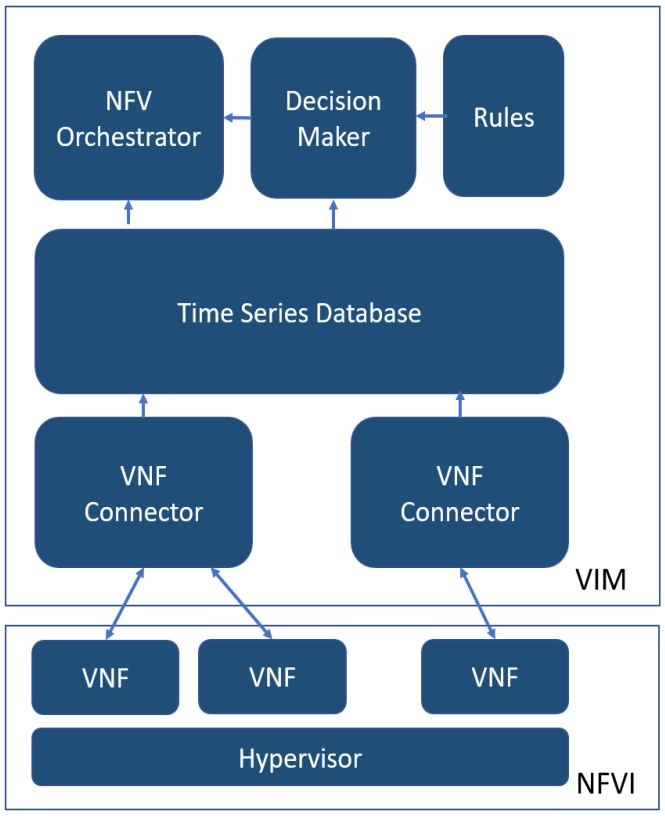

Fig.1 Integrated Monitoring Framework for Time Series Based Performance Matrix Collection

1.2 VNF Scaling

NFV enables to dynamically modify the capacity of Network Services (NSs) to face changes such as the number of users and performance variations. This can be either by increasing resources of same VM by adding/reducing CPU, RAM, Storage known as scale-up and scale-down Or by adding/removing new VMs known as scale-out and scale-in. This research focuses mainly on the scale-out and scale-in decision based scenarios.

The typical approach found in most monitoring systems and is based on generating an alarm when a metric has crossed a pre-defined threshold. The challenge is to study and implement dynamic methods for fault detection in VNF. Such methods should be based on statistical methods and self-learning approaches, detecting outliers in system behaviour and triggering alarms reactively or better proactively i.e. before the fault has to be occurred. This fault detection procedures can clearly help from the fact that the monitored services are composed of VNFs rather than generic Virtual Machines (VM). As virtual machines dedicated to traffic processing, VNFs are expected to have some common properties (e.g. the CPU load is expected to proportionally increase, not necessarily linearly, with the increase of traffic). A noticeable deviation from such correlation could, for example, indicate a potential malfunction.

The NFV scaling should be accurate at the time as well as count of instances to be scaled, aiming to avoid the unnecessary steps of allocating and releasing of virtual resources; however, to achieve a high accuracy is a non-trivial task [4]

1.3 Challenges in Scaling

A. Reliability

A high service availability is very crucial to the telco providers for reliability of their services. High availability needs a low Service Level Agreement (SLA) violation. A delay of more than appropriate scaling time may bring about SLA violations.

B. Turn Around Time

Most of the researches have consider the VM startup delay as 10 minutes. . If system is experiencing a heavy load, then the method of immediate scaling won't give smooth response time. So, it needs to have a mechanism which does this scaling activity well in advance, by considering this delay. [11]

C. Stability

Oscillation of VM count reflects flexibility of the system. The system should optimize resource utilization by adding new VM instances or removing idle VM in time. But continuous scaling actions increases the system overhead and should be reduced. It's a tradeoff between reliability and stability to provide NFV services.

D. Optimal Resource Utilization

Computing resource and network resource are of main components in the Telecommunications Cloud computing platforms. According to recent authoritative reports [20], the resource utilization ratio of modern data centers is around 15%~30%, which is a lot of waste in physical resource.

2. SCALING APPROACHES

In the literature, several approaches are proposed for scaling mechanisms that differ in the utilized technique (e.g., reactive, adaptive and predictive) [6]. Some of these such as predictive approaches are machine learning-based

2.1 Static Threshold Based (Reactive)

The static threshold based rules are very common in cloud service providers like Amazon EC2 and IBM Cloud. Pre-defined upper threshold and lower threshold are compared for each performance metric. If the performance metric is above the upper threshold at given time, scale-out action will be triggered. Whereas, scale-in decision will be made if performance matric is below lower thresholds. After each scaling action, a cool down period is applied in which scaling actions are not performed on that VNF

Selection of proper threshold values is crucial factor in the success of static threshold based approach. A lower threshold value results in frequent changes in VM configuration and a very high value reduces the responsiveness of VM to adaption of the new resource requirement. There are several techniques to find out optimum threshold values such as history based, mathematical model [12]

Static threshold based method is easy and simple to implement. However, this kind of technique is reactive, which can cause SLA (Service Level Agreement) violations. Moreover, the thresholds is difficult to choose and may needed to be frequently changed.

2.2 Dynamic Threshold Based (Adaptive)

The thresholds are initially set and automatically updated with the SLA violation. The dynamically changed thresholds are achieved by estimating the statistics and distribution of CPU utilization for physical servers. Although the idea of using a controller to handle the auto-scaling operations is appealing, it is challenging to create a reliable model

The system should be capable of automatically adjusting its threshold values depending on the workload patterns that are observed for that application. In [14] have proposed a novel technique for the efficient threshold based dynamic

consolidation of Virtual Machines with auto-tuning of the threshold values

In many IaaS environments there are multiple other factors like Virtual Machine turnaround time and stabilization time etc. that impact the newly started VM from boot time to request servicing time. If such factors are not considered while auto scaling, then they will have direct impact on Service Level Agreement (SLA) executions and user's response time. Therefore, these thresholds should be a function of load trends, so that VM readily available when needed. [11]

2.3 Time Series Analysis (Predictive)

This approach uses the time series database to identify patterns of VNF performance. Time series database contains the performance matrix values along time series and identified patterns can be used to predict the scaling decision in future times.

The easy approaches are classic exponential smoothing, auto-regression , moving average and auto-regressive moving average are applied to predict resource requirements or workload trends. [15]

Moreover new machines learning based scaling techniques based on neural networks can help to forecast the upcoming load as well. Hence scheduling the resource provisioning in a way that all the required resources will be deployed and actively available when the load increases. In the same way, it will scale-in the unneeded resources when the traffic load decreases. There are two types of neural network available to run the machine learning algorithms: DNN and RNN. The DNN, or also called feed forward neural network, is composed of several neural layers. Whereas the RNN (Recurring Neural Network) keeps state memory of the last passed activation events in the network as temporal contextual information [8]

In [14] presents a novel PRedictive Elastic reSource Scaling (PRESS) mechanism for cloud systems. PRESS extracts fine-grained dynamic patterns. It offers the forecast based on historic patterns, states or the usual arithmetic methods such as min, max and regression.

In [13] using simulation some representative auto-scaling techniques are been tested and compared them in terms of overall cost and SLO violations. It also proposes a method based on rules with dynamic thresholds that improves the performance of simple, static threshold based rules. The main conclusion provided is that any auto-scaling method is heavily dependent on the parameter tuning.

3. PROPOSED APPROACH FOR AUTO-SCALING

Proposed solution is combination of adaptive and machine learning approaches. Machine learning algorithms will be trained on past data to identify the boundary criterion for classification of matric values into scaling actions such as scale-in, scale-out or no-action. This trained algorithm then can be used to classify the real-time matrix values into scaling actions and will be notified to orchestrator. Whereas the feedback will be provided by decision maker back to time series database for inaccurate predictions to improve results.

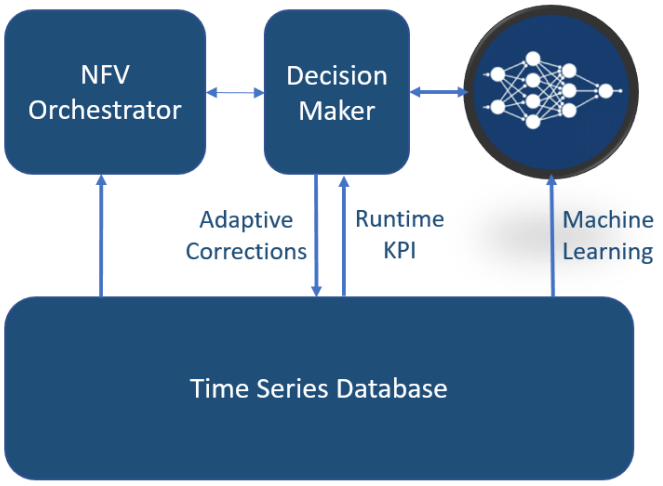

Fig.2 Combination of Adaptive & ML approach

4. CONCLUSION AND FUTURE WORK

Network Functions Virtualization brings flexibility in deployment of Network Services also allowing these services to be scaled depending on their workload demand or performance variations. There are multiple reactive and proactive approaches available as defined in literature survey. Proactive approaches such as Adaptive (Parameter Tuning) and Predictive (Predicting load in near future) are more effective compared to Reactive approaches. Machine learning based algorithms are being used in Predictive approaches.

As future work, we are proposing novel approach of parameter tuning by combination of both feedback based adaptive approach and machine learning based prediction approaches. We intend to implement this approach and compare the results with existing approaches.

REFERENCES

[1] ETSI GS NFV."Virtual Network Function Architectural Framework", V1.1.1.

[2] ETSI-GS-NFV-MAN. "Network functions virtualization (nfv): management and orchestration" (2014)

[3] Georgios Gardikis, Ioannis Koutras, George Mavroudis et al., "An integrating framework for efficient NFV monitoring", NetSoft, pp. 1-5, 2016.

[4] C. H. T. Arteaga, F. Rissoi and O. M. C. Rendon, "An adaptive scaling mechanism for managing performance variations in network functions virtualization: A case study in an NFV-based EPC," 2017 13th International Conference on Network and Service Management (CNSM), Tokyo, 2017.

[5] B. Han V. Gopalakrishnan L. ji S. Lee "Network functions virtualization: Challenges and opportunities for innovations

[6] P. Tang F. Li W. Zhou W. Hu L. Yang "Efficient auto-scaling approach in the telco cloud using self-learning algorithm" IEEE Global Communications Conference (GLOBECOM) 2015

[7] Y.-J. Chang et al., "Scalable and Elastic Telecommunication Services in the Cloud", Bell Labs Technical Journal, vol. 17, no. 2, pp. 81-96, 2012.

[8] Imad ALAWE "Smart scaling of the 5G core network", http://www.eurecom.fr/en/publication/5648/download/comsys-publi-5648_1.pdf

[9] Carella, Giuseppe Antonio. "An extensible and customizable framework for the management and orchestration of emerging software-based networks." (2018)

[10] Lorido-Botrán, Tania & Miguel-Alonso, Jose & Antonio Lozano, Jose. (2014). "A Review of Auto-scaling Techniques for Elastic Applications in Cloud Environments". Journal of Grid Computing. 12. 10.1007/s10723-014-9314-7

[11] Kartheek Kanagala, K. Chandra Sekaran. "An approach for dynamic scaling of resource in enterprise cloud." 2013 IEEE International Conference on Cloud Computing Technology and Science. DOI 10.1109/ CloudCom.2013.167

[12] M.K. Mohan Murthy, H.A. Sanjay, Jumnal Anand. "Threshold based auto scaling of virtual machines in cloud environment". NPC 2014, LNCS 8707, pp. 247–256, 2014

[13] Lorido-Botran, T., Miguel-Alonso, J., Lozano, J. A.: "Comparison of Auto-scaling Techniques for Cloud Environments". In: Alberto, A., Del Barrio, G.B. (eds.) Actas de las XXIV Jornadas de Paralelismo, Servicio de Publicaciones (2013)

[14] A Beloglazov and R Buyya. "Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers". In Proceedings of the 8th International Workshop on Middleware for Grids, Clouds and e-Science, page 4. ACM, 2010

[15] Z. Gong , X. Gu, J. Wilkes "Press: predictive elastic resource scaling for cloud systems", In: 2010 International Conference on Network and Service Management (CNSM), pp. 9–16 (2010)