JOINT SCALING OF VIRTUAL NETWORK FUNCTIONS AND NETWORK TO MINIMIZE OPERATIONAL AND LEASING COST

Sabidur Rahman
Friday Group Meeting, Netlab
UC Davis

Agenda

- Motivation
- Problem statement
- Progress and update
- Future works

Auto-scaling (1)

“Autoscaling, also spelled **auto scaling** or **auto-scaling**, is a method used in cloud computing, whereby the amount of computational resources in a server farm, typically measured in terms of the number of active servers, scales automatically based on the load on the farm.”

Amazon Web Services (AWS)
Netflix
Microsoft's Windows Azure
Google Cloud Platform
Facebook

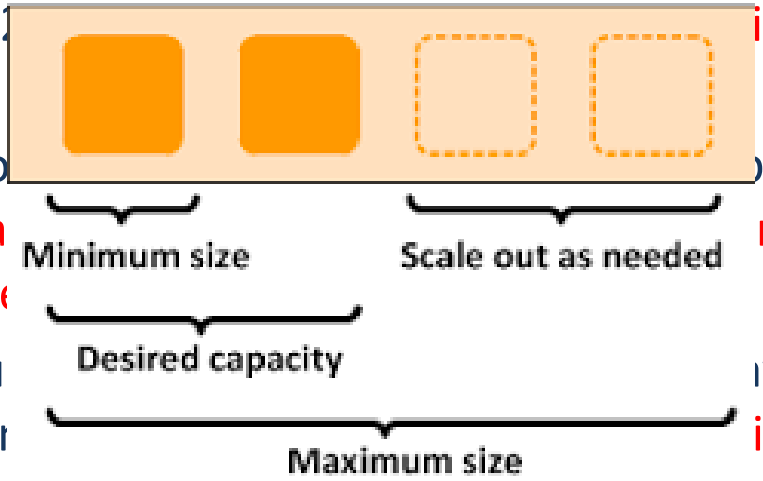
1. https://en.wikipedia.org/wiki/Autoscaling

Auto-scaling (2)

"Auto Scaling helps you maintain application availability and allows you to scale your Amazon EC2 ically according to conditions you define.

….Auto Scaling can also ber of Amazon EC2 instances during dema nce and decrease capacity during lulls to re

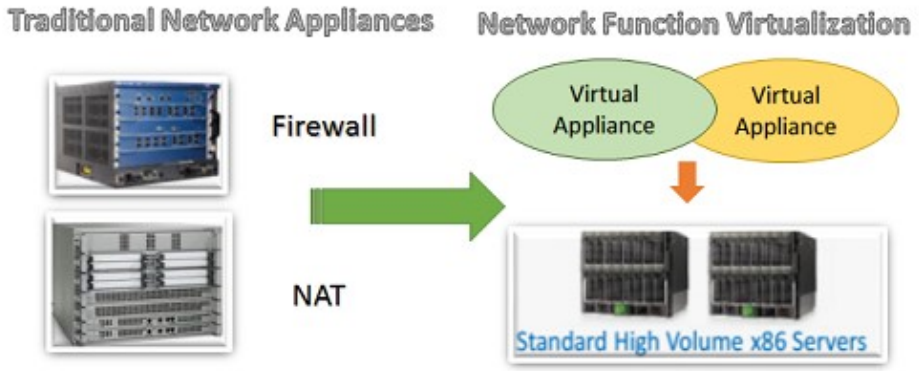
Auto Scaling is well su ve stable demand patterns or that experien ity in usage."



2. https://aws.amazon.com/autoscaling/

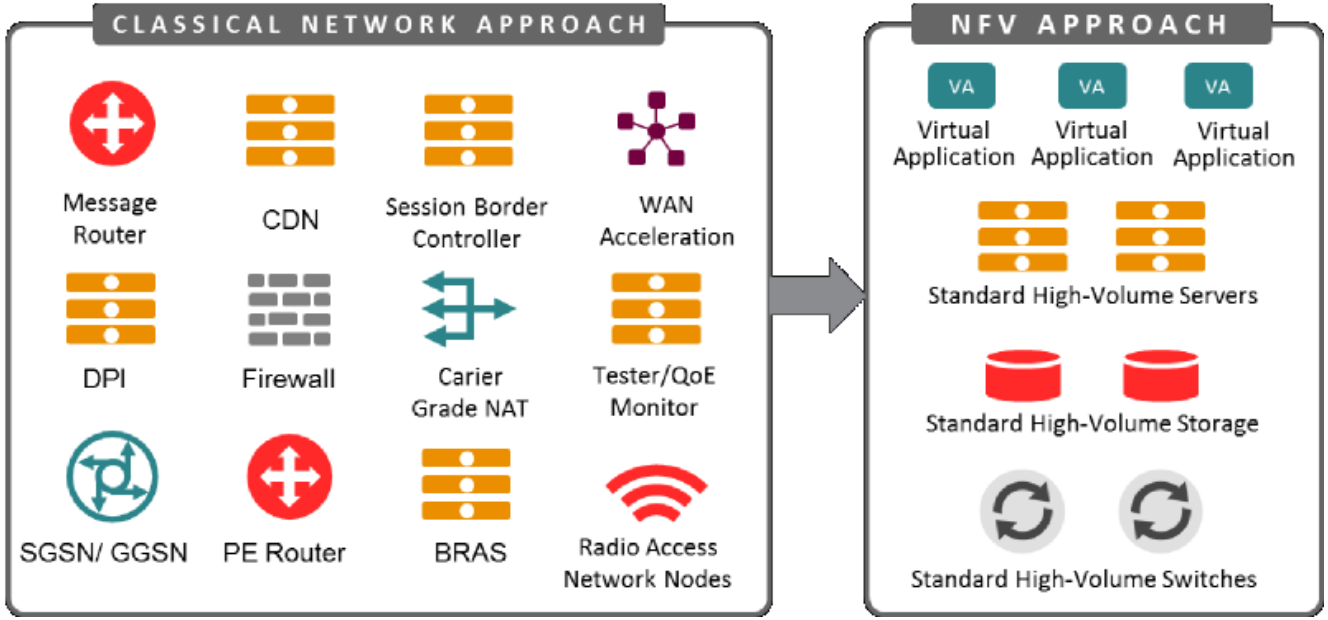
Auto-scaling of network resources

- Broadband Network Gateways (BNGs)
- Evolved Packet Core (EPC)
- Firewalls
- Deep Packet Inspection (DPI)
- Data exfiltration systems
- NATs
- Web Proxies
- Load balancers
- Content caching
- Parental control



3. Palkar S, Lan C, Han S, Jang K, Panda A, Ratnasamy S, Rizzo L, Shenker S. E2: a framework for NFV applications. InProceedings of the 25th Symposium on Operating Systems Principles 2015 Oct 4 (pp. 121-136). ACM.

Network Function virtualization



4. http://www.alepo.com/white-papers/alepo-in-the-virtualized-core-network/

5. Gupta A, Habib MF, Chowdhury P, Tornatore M, Mukherjee B. Joint virtual network function placement and routing of traffic in operator networks. UC Davis, Davis, CA, USA, Tech. Rep. 2015 Apr 20.

Motivation

- Content Distribution Networks (CDNs) [10]: Netflix, Akamai.
- Telecom networks [11]: AT&T, Verizon.
- Data Center Networks [13]: Google, Amazon, Facebook.
- Mobile Virtual Network Operators [12]: Boost Mobile (Sprint), Cricket Wireless (AT&T), MetroPCS (T-Mobile US)
- Software-defined Data Center [14]
- Network function outsourcing

10. Mandal U, Chowdhury P, Lange C, Gladisch A, Mukherjee B. Energy-efficient networking for content distribution over telecom network infrastructure. Optical Switching and Networking. 2013 Nov 30;10(4):393-405.
11. Zhang Y, Chowdhury P, Tornatore M, Mukherjee B. Energy efficiency in telecom optical networks. IEEE Communications Surveys & Tutorials. 2010 Oct ;12(4):441-58.
12. Zarinni F, Chakraborty A, Sekar V, Das SR, Gill P. A first look at performance in mobile virtual network operators. InProceedings of the 2014 Conference on Internet Measurement Conference 2014 Nov 5 (pp. 165-172). ACM.
13. Heller B, Seetharaman S, Mahadevan P, Yiakoumis Y, Sharma P, Banerjee S, McKeown N. ElasticTree: Saving Energy in Data Center Networks. InNSDI 2010 Apr 28 (Vol. 10, pp. 249-264).
14. http://www.vmware.com/solutions/software-defined-datacenter.html?src=phd709

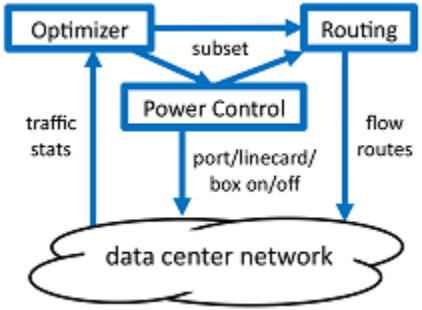
Literature review (1)

10. Mandal U, Chowdhury P, Lange C, Gladisch A, Mukherjee B. Energy-efficient networking for content distribution over telecom network infrastructure. Optical Switching and Networking. 2013 Nov 30;10(4):393-405.

- **Focus**: Content distribution over telecom network
- Energy consumption model, analysis and content-placement techniques to reduce energy cost
- Storage power consumption and transmission power consumption
- Time-varying traffic irregularities
- More content replicas during peak load and less replicas during off-peak load

Literature review (2)

13. Heller B, Seetharaman S, Mahadevan P, Yiakoumis Y, Sharma P, Banerjee S, McKeown N. ElasticTree: Saving Energy in Data Center Networks. In NSDI 2010 Apr 28 (Vol. 10, pp. 249-264).

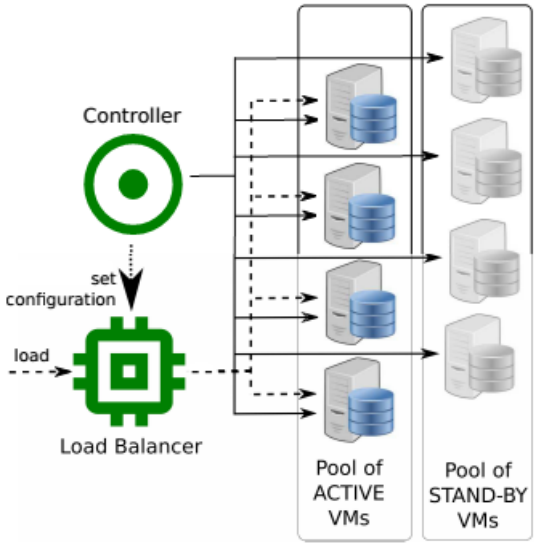


- **Focus**: Data center networks
- Scale up and down to save energy
- Dynamically adjust link and switches to satisfy changing traffic load
- Optimizer monitors traffic to choose set of elements needed to meet performance and fault tolerance goals.
- Formal model, Greedy bin-packer, topology-aware heuristic and demand prediction-based method

Literature (3)

15. Avresky DR, Di Sanzo P, Pellegrini A, Ciciani B, Forte L. Proactive Scalability and Management of Resources in Hybrid Clouds via Machine Learning. In Network Computing and Applications (NCA), 2015 IEEE 14th International Symposium on 2015 Sep 28 (pp. 114-119). IEEE.

- A proactive system scale up / scale down technique
- Machine learning models for predicting failures caused by accumulation of anomalies (Software/Hardware)
- When a VM joins (or leaves) a region, the region workload is automatically spread across local VMs



Phung-Duc T, Ren Y, Chen JC, Yu ZW. Design and Analysis of Deadline and Budget Constrained Autoscaling (DBCA) Algorithm for 5G Mobile Networks. arXiv preprint arXiv:1609.09368. 2016 Sep 29.

- VNFs can be dynamically scale-in/out to meet the performance desire
- Auto-scaling algorithm for desired characteristics with low operation cost and low latency
- Tradeoff between performance and operation cost
- NFV enabled Evolved Packet Core (EPC) is modeled as queueing model
- Legacy network equipment are considered as **reserved a block of servers**
- VNF instances are powered on and off according to the number of job requests present.

Tang P, Li F, Zhou W, Hu W, Yang L. Efficient Auto-Scaling Approach in the Telco Cloud Using Self-Learning Algorithm. In 2015 IEEE Global Communications Conference (GLOBECOM) 2015 Dec 6 (pp. 1-6). IEEE.

- Provision and orchestration of **physical and virtual resource** is crucial for both Quality of Service (QoS) guarantee and cost management in cloud computing environment.
- SLA-aware and Resource-efficient Self-learning Approach (SRSA) for auto-scaling policy decision
- Busy-and-idle scenario and burst-traffic scenario

Tang P, Li F, Zhou W, Hu W, Yang L. Efficient Auto-Scaling Approach in the Telco Cloud Using Self-Learning Algorithm. In 2015 IEEE Global Communications Conference (GLOBECOM) 2015 Dec 6 (pp. 1-6). IEEE.

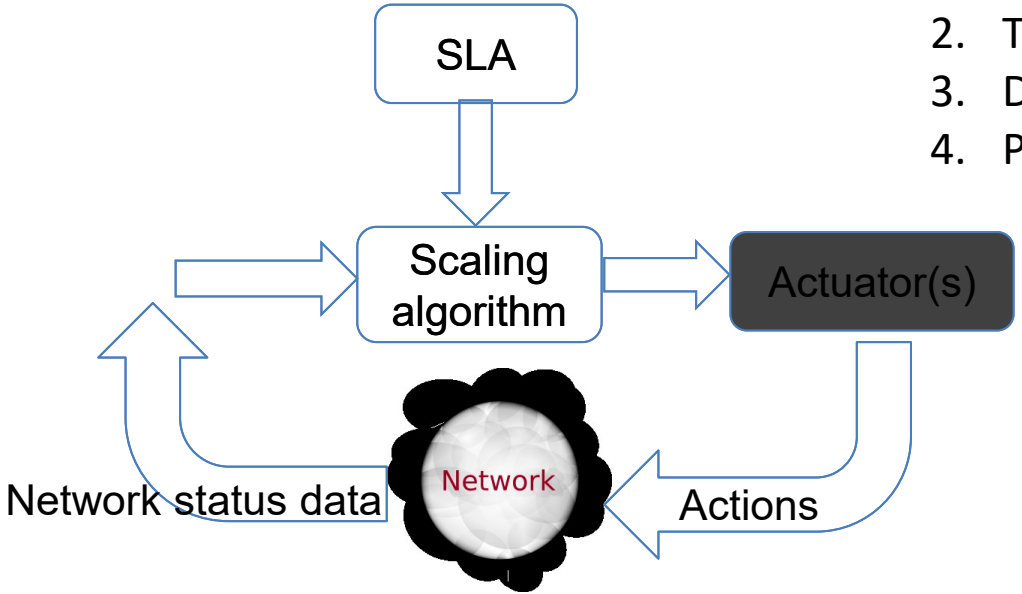
- Provision and orchestration of **physical and virtual resource** is crucial for both Quality of Service (QoS) guarantee and cost management in cloud computing environment.
- SLA-aware and Resource-efficient Self-learning Approach (SRSA) for auto-scaling policy decision
- Busy-and-idle scenario and burst-traffic scenario

Problem statement

Given: Network topology, network traffic data, SLA

Objective: Joint scaling of VNFs and network to minimize network operation cost (and network leasing cost).

Problem overview

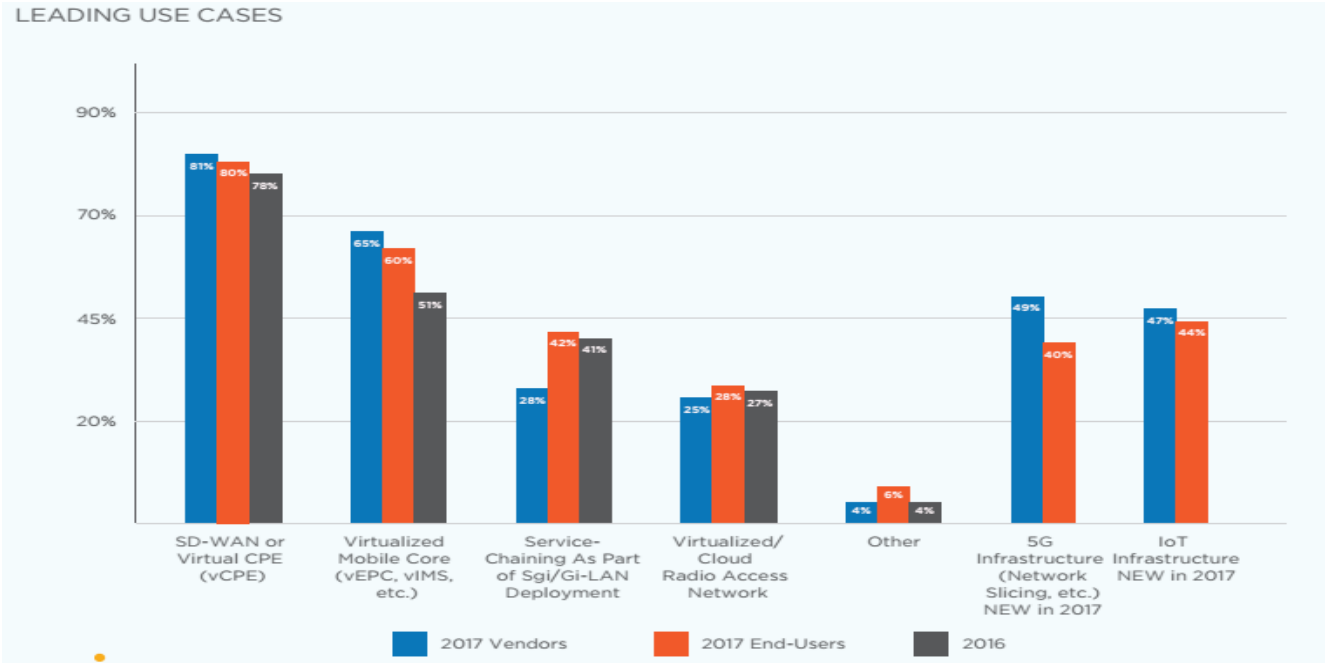


Unsolved points from last meeting:

1. A scenario with larger network.
2. Traffic trace/generation.
3. Design of machine learning model.
4. Performance of method.

Usecase for NFV

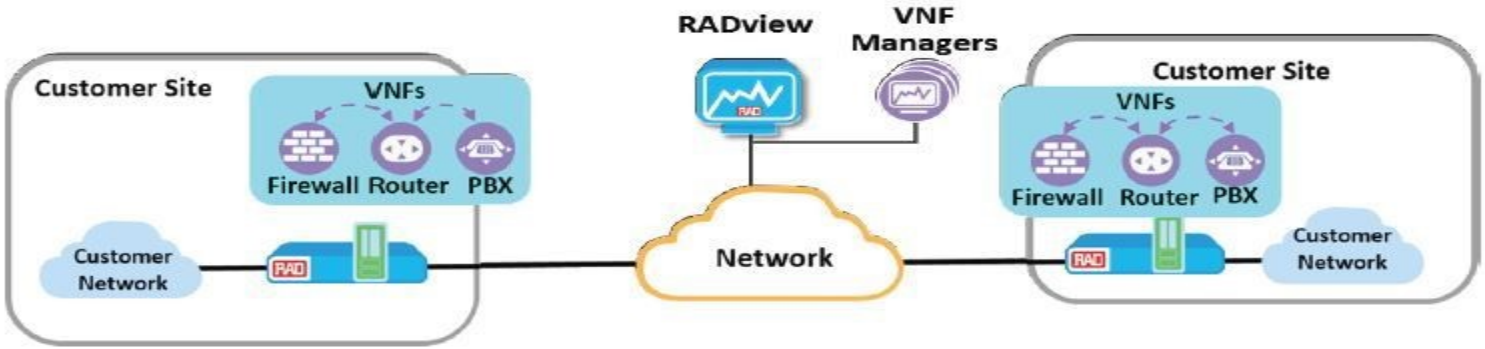
vCPE environments (80%)
virtualized mobile cores (60%)



SDxCentral NFV survey 2017

UCDAVIS

SD-WAN and VNFs

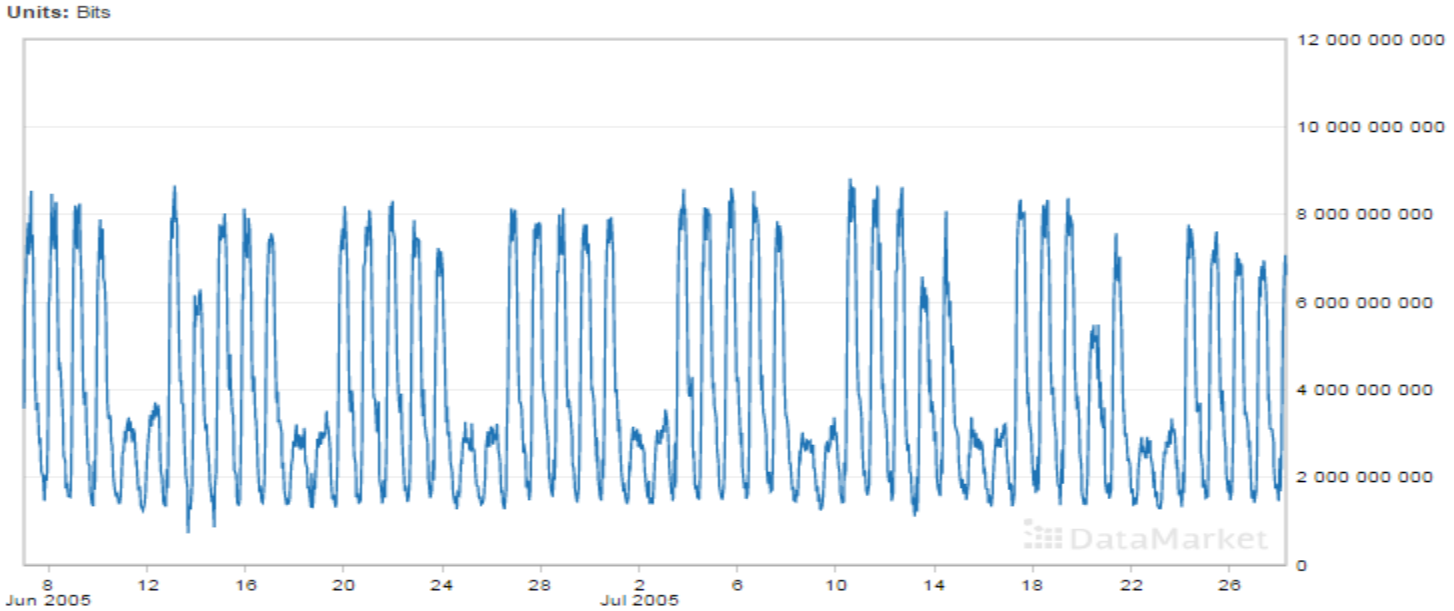


http://cietelematica.it/eng/featured-products/rad--intel---maximizing-vcpe-flexibility/

SD-WAN and VNFs

*Branches and HQ/DC would later be placed on a backbone network.

Data Source



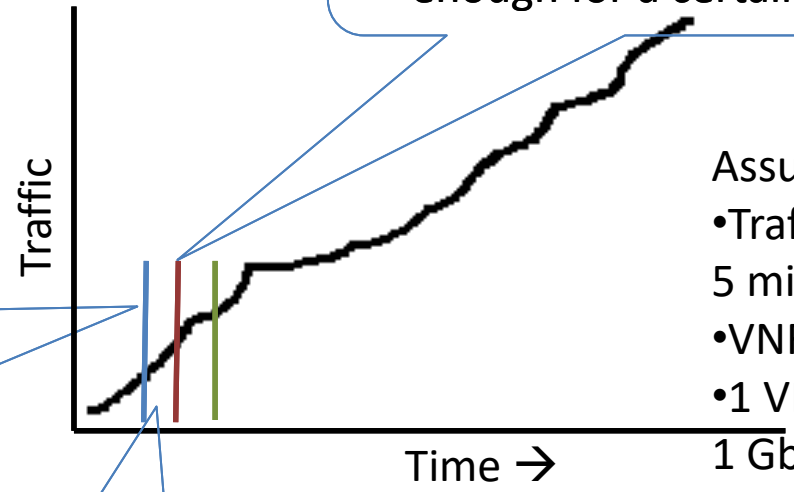
The data corresponds to a transatlantic link and was collected from 06:57 hours on 7 June to 11:17 hours on 31 July 2005. Data collected at five minute intervals.

Source: Time Series Data Library (citing: P. Cortez, M. Rio, M. Rocha and P. Sousa. Multiscale Internet Traffic Forecasting using Neural Networks and Time Series Methods. In Expert Systems, Wiley-Blackwell, In press.)

UCDAVIS

From data to answer?

But, it is not trivial to answer if the same number of VNFs are enough for a certain period.



At a given point, it is very trivial to calculate the necessary number of VNFs.

5 min period

Assumptions:
- Traffic measurement granularity 5 min
- VNF scaling granularity 10 min.
- 1 VNF instance required to serve 1 Gb traffic

From data to answer

3.562279127,3.710215571,3.877469703,3.876354871,4.582542581,5.016336869,5.202513642,5.410604985,5.408071320…….

```
@ATTRIBUTE time REAL
@ATTRIBUTE traffic REAL
@ATTRIBUTE delta    REAL
@ATTRIBUTE class    {1,2,3,4,5,6,7,8,9,10}

@DATA
0,3.710215571,0.14793644399999994,4
1,3.877469703,0.16725413200000006,5
2,3.876354871,-0.0011148319999998435,6
3,4.582542581,0.70618771,6
4,5.016336869,0.43379428799999964,6
```

3.562279127,3.710215571,3.877469703,3.876354871

3.710215571,3.877469703,3.876354871,4.582542581

Would more features help?

Machine learning

No extra VNFs, no SLA violation

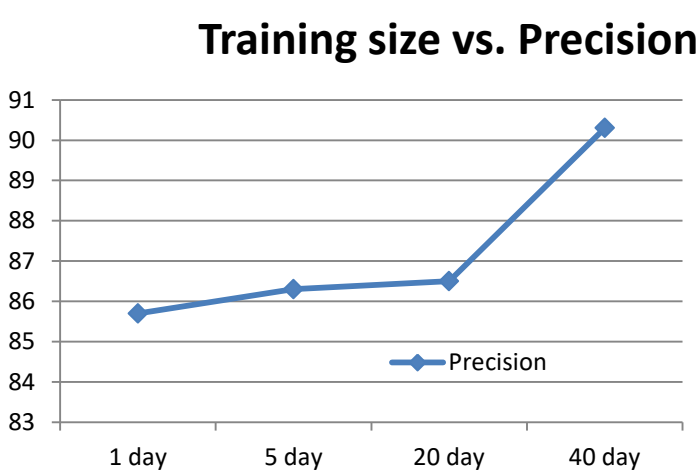
Model design:

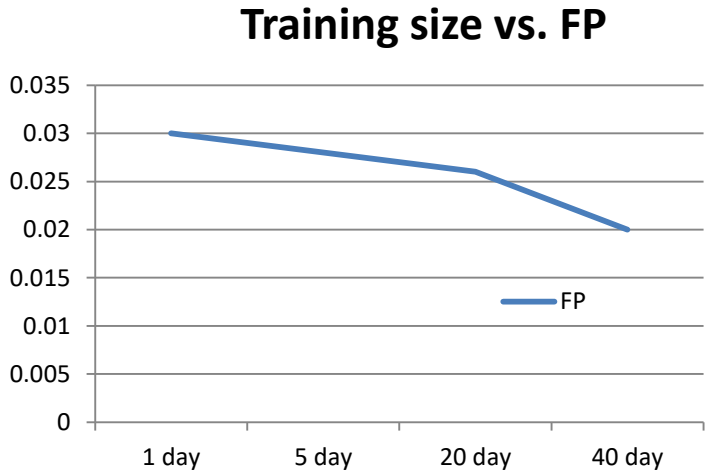
- Use an offline method to calculate ground truth.
- For a given time and traffic: “how many VNFs should the premise have to maintain SLA requirement?”
- (time, traffic, traffic change) -> number of VNFs required for the period

Evaluation and improvement:

- Can I train with 5 days of results and predict the 6th day’s result?
- How close is the prediction to actual value?
- Does the model improve with more training data? (1day=85.7,40day = 90.9 using Random tree)

Results: performance vs. sample size





Results: Algorithms compared

Algorithm	Training time (s)	Test time (s)	Precision	ROC Area	FP
Random Tree	0.09	0.01	90.9	94.1	0.02
Random Forest	4.91	0.16	93.8	99.6	0.013
Bayesian Network	0.4	0.03	92.5	99.8	0.018
Neural Network	41.11	0.03	94.0	99.8	0.009

*40 day training data

Work in progress

-How often should we scale the VNFs? (Every 1 min? 5 min?)

- Scaling too often will make VNF management too unstable, in case of traffic that changes too often.
- Scaling less often will lead to SLA violation or excess energy consumption.

-Can machine learning/data tell when to scale? (What time in future we need extra VNFs and how many?)

-Traffic trace should be more dynamic. Traffic may change a lot during 5 min period.

-Translating the machine learning result into cost of operation and cost of leasing.


Thank You