

AppSlice: A system for application-centric design of 5G and edge computing applications

Murugan Sankaradas
NEC Laboratories America
 Princeton, NJ
 murugs@nec-labs.com

Kunal Rao
NEC Laboratories America
 Princeton, NJ
 kunal@nec-labs.com

Srimat Chakradhar
NEC Laboratories America
 Princeton, NJ
 chak@nec-labs.com

Abstract—Applications that use edge computing and 5G to improve response times consume both compute and network resources. However, 5G networks manage only network resources without considering the application’s compute requirements, and container orchestration frameworks manage only compute resources without considering the application’s network requirements. We observe that there is a complex coupling between an application’s compute and network usage, which can be leveraged to improve application performance and resource utilization. We propose a new, declarative abstraction called *app slice* that jointly considers the application’s compute and network requirements. This abstraction leverages container management systems to manage edge computing resources, and 5G network stacks to manage network resources, while the joint consideration of coupling between compute and network usage is explicitly managed by a new runtime system, which delivers the declarative semantics of the app slice. The runtime system also jointly manages the edge compute and network resource usage automatically across different edge computing environments and 5G networks by using two adaptive algorithms. We implement a complex, real-world, real-time monitoring application using the proposed app slice abstraction, and demonstrate on a private 5G/LTE testbed that the proposed runtime system significantly improves the application performance and resource usage when compared with the case where the coupling between the compute and network resource usage is ignored.

Index Terms—5G, edge computing, slicing, application-centric design, specification, runtime

I. INTRODUCTION

Edge computing and 5G are inextricably linked technologies that promise to enable huge amounts of data to be processed in real-time and to significantly improve the response times of low-latency applications. Edge computing [1] brings compute, storage, switching and control functions relatively close to end users and IoT endpoints. Emerging tiers of network and compute is shown in Figure 1. In these tiers, critical data can be processed at the edge of the network, while less urgent data can be sent to the cloud for data processing.

5G networks promise ultra-low latency, high reliability, high bandwidth and high device density, and they are expected to enable a wide variety of emerging applications like remote operation, remote maintenance, augmented reality, mobile workforce and enterprise applications (like payments, tactile, V2X [2] and real-time surveillance). Often, these applications require low latency (0.5 to 10 milliseconds), high data rate

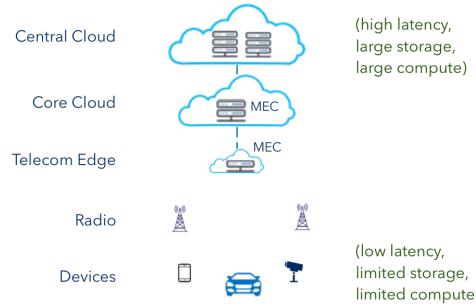


Figure 1: Compute and network tiers

(10 to 1000 Mbps), and high device density (1000s of sensing devices).

Although it is widely recognized that both edge computing and 5G networks are required to realize ultra low-latency applications, there is no principled, application-centric approach that automatically, and dynamically, optimizes the application across a range of edge computing platforms and 5G network stacks [3] [4] [5]. Today, 5G networking vendors manage and provision network resources (soft or hard network slicing [6] [7]) without considering the application’s compute requirements [8] [9]. Similarly, container orchestration frameworks like Kubernetes [10] manage and provision compute resources without considering the application’s network requirements. It is also the case that these container management systems manage or provision compute resources only within a specific tier in the layered computing fabric shown in Figure 1, rather than across all tiers of the computing fabric.

In order to have a unified, application-centric view that jointly considers both compute and network resources for an application across diverse edge computing environments and 5G network stacks, we propose a new abstraction layer called *app slice*, which jointly considers the application’s compute as well as the network resources. This abstraction layer leverages container management systems to manage edge computing resources, and 5G network stacks to manage network resources, while the joint consideration of coupling between compute and network resource usage is explicitly managed by a new runtime for the proposed app slice abstraction.

We make the following key contributions in this paper.

- We propose a new, application-centric declarative specification, called *app slice*, which allows joint specification

of compute and network requirements of the application as a whole, as well as the individual functions (or microservices) that make up the application.

- We propose a new runtime system, which realizes the declarative semantics in the app slice specification, and jointly manages the edge compute and network resources across different edge computing environments and 5G networks by using two adaptive algorithms.
- We implement a complex, real-world, real-time monitoring application using the proposed app slice abstraction, and demonstrate on a private 5G/LTE testbed that the proposed app slice specification and runtime system significantly improve the application performance and resource usage when compared with the case where the complex coupling between the compute and network resource usage is ignored.

II. DECLARATIVE, APP SLICE SPECIFICATION

We model the application as a set of functions or microservices, specified through an *app specification*. App specification includes application details, the functions that comprise the entire application, function details along with the instances of functions and finally the inter-connection between various function instances.

The proposed declarative *app slice specification* consists of two parts: an application-level specification that captures application-level requirements like latency and bandwidth, and a function-level specification that captures the compute and network requirements of each function.

A. Application-level specification

It captures the desired application requirements:

- *latency*: End-to-end application latency (in milliseconds).
- *bandwidth*: Overall network bandwidth (in Mbps).
- *deviceCount*: Total number of devices the application connects to, or expects to receive data streams from.
- *reliability*: Desired reliability (between 0 and 1). 0 being unreliable and 1 being totally reliable.

B. Function-level specification

It includes compute and network specification.

1) *Network specification*: The network requirements of a function are specified as part of the this specification.

- *latency*: Maximum tolerable network latency (in milliseconds). When function A's output is fed to another function B as input, then latency specification for function B is the latency of the link connecting function A to function B.
- *throughputGBR*: Guaranteed network bandwidth (in Mbps) required by the function.
- *throughputMBR*: Maximum bandwidth (in Mbps) that can be consumed by the function.
- *packetErrorRate*: Ratio of the number of incorrectly received packets and the total number of received packets.
- *duration*: Optional duration (in milliseconds) for which the network guarantees should be provided for the function. Default value for this is “auto”, indicating that the runtime can choose and decide this value dynamically.

2) *Compute specification*: Function's compute requirements are specified as part of this specification.

- *minCPUCores*: Desired minimum CPU resources (absolute cpu units). 1 represents either 1 vCPU/core on the cloud or 1 hyperthread on bare-metal Intel processors. 1 cpu unit is divided into 1000 “millicpus” and the finest granularity that can be specified is “1m” (1 millicpu).
- *maxCPUCores*: Maximum CPU cores (absolute cpu units) that the function can use.
- *minMemory*: Desired minimum memory (in bytes).
- *maxMemory*: Maximum allowable memory (in bytes).
- *tier*: Optional parameter to specify specific tier i.e. “device”, “edge” or “cloud” in the computing fabric where the function should run. Default value for this is “auto”, indicating that the function can run anywhere in the computing fabric.

III. APP SLICE RUNTIME SYSTEM (RS)

Our RS, shown in Figure 2, (a) is integrated with the application itself (b) sits on top of the compute and 5G network infrastructure (c) relies on standard APIs from the underlying 5G network for network slices (d) relies on standard APIs from underlying compute infrastructure like Kubernetes [10] for compute slices and (e) manages the overall execution of the application. RS consists of Resource Manager, App Slice Controller and App Slice Monitor.

A. Resource Manager (RM)

RM manages execution of total application by co-ordinating with ASC and ASM. For resource allocation, RM first checks application level slice specifications. Then, for each function in the application, RM follows the algorithm shown in Algorithm 1. First priority is given to meeting function's compute and network requirement and second priority is given to the cost. If the resource request cannot be met for the application and all its associated functions, then RM reports it to the application, and leaves it to the application and associated functions to take appropriate actions.

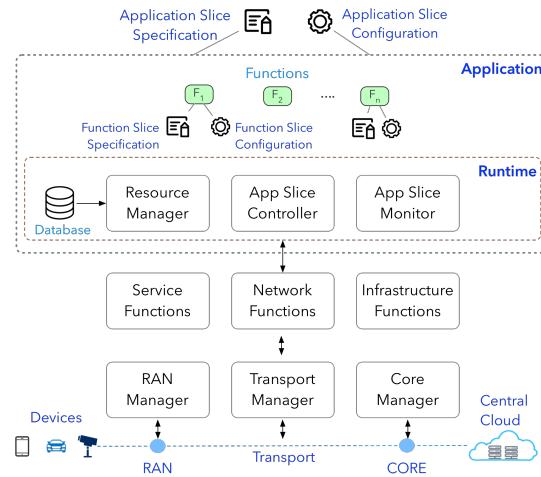


Figure 2: App slice runtime

Algorithm 1 RM resource allocation

Input: Resource request per function (for application)
Output: Allocated Resources per function (for application)

```

1:  $\triangleright$  key: function, value: allocated tier resources
2: Initialize map  $\rightarrow$  resources
3: for function  $\in$  functions do
4:    $\triangleright$  Order tiers in ascending order of resource cost
5:    $\triangleright$  Cheaper tiers are checked before expensive ones
6:   for tier  $\in$  tiers do
7:      $\triangleright$  Match requested with available tier resources
8:     if matchResources(c_r, n_r, tc_r, tn_r) then
9:        $\triangleright$  Allocate tier resources to function
10:      resources[app]  $\leftarrow$  tc_r, tn_r;
11:      break;
12:    end if
13:   end for
14: end for
15:  $\triangleright$  Return allocated resources
16: return resources

```

Algorithm 2 RM dynamic resource adjustment

```

1: while true do
2:   for function  $\in$  functions do
3:      $\triangleright$  check if resource conditions have changed
4:     if resourceConditionChanged(c_r, n_r) then
5:        $\triangleright$  check if new resources are available
6:       resources  $\leftarrow$  getResources(c_r, n_r)
7:       if resources then
8:          $\triangleright$  schedule on new resources
9:         scheduleFunction(function, resources)
10:      else
11:         $\triangleright$  report error for the function
12:        reportError(function)
13:      end if
14:    end if
15:   end for
16:   sleep(interval)
17: end while

```

As various functions continue to run, RM periodically monitors the status of these functions and adjusts the resources, if needed. To do so, RM follows Algorithm 2, where at every *interval* seconds, which is configurable (in our experiments we set this interval to be 2 seconds), RM checks across all the functions that are executing. Specifically, RM checks if the resource requirements of a function are being met or not by the allocated tier's compute and network resources. If for whatever reason (change in operating conditions/input content, load burst, network disruption or hardware failure) the network or compute resources are found to be insufficient, then RM tries to find additional resources. Along with checking if additional resources are needed, RM also checks if resources are under-utilized, and if so, releases them.

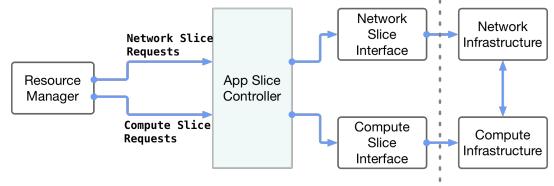


Figure 3: App slice Controller

B. App Slice Controller (ASC)

ASC, shown in Fig. 3, follows directions from RM and manages the slicing, including compute and network slicing for functions. Although the name says “controller”, ASC does not actually “control” low-level network or compute slices, rather it uses the underlying infrastructure to manage network and compute slices for various functions. For network slices, to enforce admission control, we create a custom layer on top of existing network vendors [11] and expose this layer to ASC. For compute slices, ASC directly interacts with the underlying compute infrastructure.

C. App Slice Monitor (ASM)

ASM continuously monitors and collects various compute and network usage metrics. Specifically, in our setup (described in Section IV-A), we have 5G network (and network slicing) between devices (CPE) and edge (MEC), and ASM uses the underlying 5G network’s APIs to measure latency, throughput and packet error rate. Beyond MEC and into the cloud, there is no 5G network (and therefore no network slicing) as it is over WAN. For this, ASM uses tools like iPerf3 [12] to collect network related metrics. With respect to collecting compute related metrics, ASM uses underlying compute infrastructure’s APIs, in our setup Kubernetes’ APIs. Note that unlike network slicing, compute slicing extends all the way to the cloud and there is Kubernetes cluster setup at each tier (devices, MEC and cloud). ASM communicates with appropriate Kubernetes cluster in the specific tier to collect compute related metrics. These metrics are used by RM to manage application execution.

IV. EXPERIMENTAL SETUP

A. Testbed

A high level overview of our testbed is shown in Fig. 4. In our testbed, we have used wireless gateways from Multitech [13] to connect Customer Premise Equipment (CPE) (cameras and video servers) over private 5G to Access Points from Celona [11]. Data and control plane traffic from the

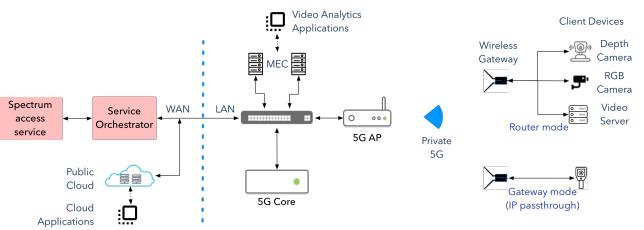


Figure 4: Architecture of testbed

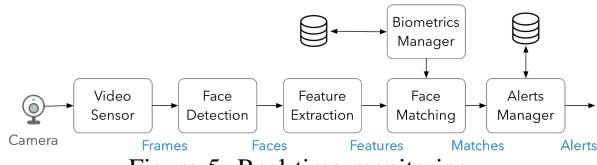


Figure 5: Real-time monitoring

access points is terminated at the Core (from Celona). Core is activated and configured remotely via Celona’s Service Orchestrator. Organizations can set specific SLAs and network requirements for different device groups and application types using network slicing. We have Multi-access Edge Computing (MEC) [14] servers connected in the same LAN with one master and three worker node servers. Master node is equipped with 10-core Intel core i9 CPU and the three worker nodes are equipped with 24-core Intel CPU and with NVIDIA RTX 2080 Ti GPUs.

B. Application: Real-time monitoring (RTM)

We implemented and used RTM video analytics application for our experiments. RTM provides fast and reliable identification of pre-registered individuals using face recognition technology. Various functions of this application, along with the pipeline is shown in Fig. 5. “Video Sensor” extracts frames, “Face Detection” extracts faces, “Feature Extraction” extracts unique facial features, “Face Matching” provides face matches (using pre-registered face gallery from “Biometrics Manager”), and finally these face matches are then stored and delivered as alerts through “Alerts Manager”.

C. Video streaming setup

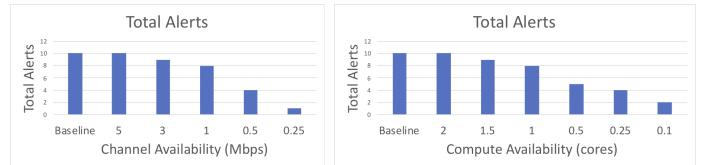
We used *Labelled movie trailer dataset (LMTD)* [15] for our experiments. We selected a video sequence containing 22 unique people from this dataset and used it as input video feed. 10 among the 22 were pre-registered in “Biometrics manager”. A combination of IP cameras (AXIS Q3515) and video streaming servers was used to generate large video traffic into our testbed. IP cameras were pointed to large display, which was looping video sequence.

V. RESULTS

In order to study the impact of compute and network resources on the accuracy of insights from the application, we have to consider application-level metrics. Therefore, in our experiments, for RTM application, we chose to measure the total number of alerts produced by the application as a measure of performance of the application. Different applications can have different metrics.

A. Performance without app slice

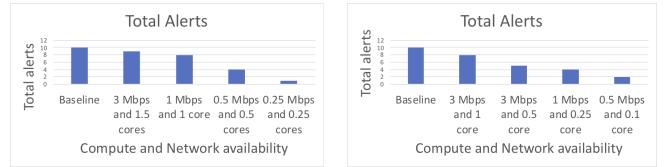
1) *Impact of varying network:* Fig. 6a shows the impact on application performance when the available network to the application varies from 5 Mbps to 0.25 Mbps (note that there is enough compute available for the application). To reduce the available network for the application, we manually pump additional traffic through the network, which impacts the application performance adversely. We see that the total number of alerts goes down from 10 to 1.



(a) Impact of varying network

(b) Impact of varying compute

Figure 6: Performance without app slice



(a) Network as bottleneck

(b) Compute as bottleneck

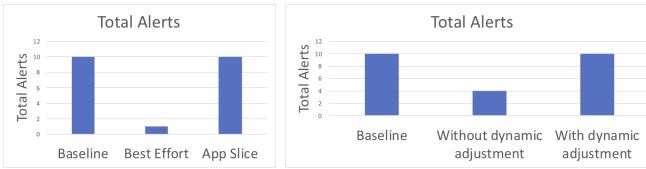
Figure 7: Impact of varying compute and network

2) *Impact of varying compute:* Fig. 6b shows the impact on application performance as the available compute goes down from 2 to 0.1 (note that there is enough network available for the application). We did this for face detection component, which is compute-intensive function and critical in determining application-level accuracy. To reduce the available cores, we used stress-ng [16], which is a CPU load generation tool. We observe that as the available compute reduces, the total number of alerts drops from 10 all the way to 1. This is because fewer frames are processed as available compute goes down (frames with face match never get processed).

3) *Impact of varying compute and network:* Next, we varied the network as well as the compute resources at the same time and observed RTM application performance. Particularly, we observed the total number of alerts when we gradually changed the network resources down from 5 Mbps to 0.25 Mbps, and compute resources down from 2 cores to 0.1 cores in various combinations. We observed that as the network and compute resources went down, the total number of alerts also went down from 10 to 1 (degradation of 90 %), as shown in Fig. 7. In Fig. 7a we see that at 0.5 Mbps and 0.5 cores, although compute was capable of receiving 5 alerts (Fig. 6b), we only saw 4 alerts, indicating that network was bottleneck. Similarly, in Fig. 7b, we observe that at 3 Mbps and 1 cores, although network was capable of receiving 9 alerts (Fig. 6a), we only saw 8 alerts, indicating that compute was bottleneck. Similar trend is observed for different combination where either network or compute can become bottleneck and have adverse effect on the application.

B. Performance with app slice

We measured the performance of RTM application, once the app slice (which includes the network and compute requirements for each function) is specified and the application is run through our app slice RS. Note that app slice RS is placed at MEC in our setup. As before, we simulated a condition where there is external load and in such condition we conducted this experiment. As shown in Fig. 8a, we see that when there is no



(a) Impact of app slice (b) Dynamic resource adjustment

Figure 8: Performance with app slice

app slice specified i.e. under “Best effort” conditions, the total number of alerts is only 1, whereas in the presence of “App slice”, the total number of alerts is 10 (same as baseline).

Fig. 8b shows the impact of dynamic resource adjustment. Here, we assume that based on the initial profiling of the application, the app-slice specified network bandwidth is 0.5 Mbps and compute is 2 cores. With this specification, and without any dynamic resource adjustment, we observed that only 4 alerts were received. However, when we use dynamic resource adjustment, our RS identifies (based on the compute and network resource usage) that network is the bottleneck and it increases the allocated network resources to 5 Mbps, while keeping the compute as 2 cores. This results in the number of received alerts increasing to 10 (which is same as the baseline).

This shows that our RS understands the coupling between network and compute, and is able to dynamically adjust appropriate resources so that application performance is optimized. Our methodology and design are applicable to a wide range of video analytics applications and we use RTM application as an illustrative example.

VI. RELATED WORK

Existing standard specifications like TOSCA [17] do apply to applications modelled as a collection of services. However, they are focused only on the cloud and do not extend to multiple tiers, like compute resources in devices, or edge (MEC). Network slicing applied to smart grid [18] [19] or healthcare [20] mainly relies only on network slice specification, which is about network and compute resources required to execute (virtualized or physical) network functions. However, applications on top of the network can use a variety of compute resources, which are not specified as part of the network slice specification.

Another recent work [21] also proposes to slice computation and communication resources. They propose vertical and horizontal slicing of the air interface, radio access network, core and the virtualized computing resources available to execute network function virtualization. However, there is no joint consideration of application’s compute and network resource usage. To the best of our knowledge, our proposal is the first to jointly consider compute and network requirements to improve efficiency and performance of 5G applications across different edge compute environments and 5G network stacks.

VII. CONCLUSION

Recognizing the complex coupling between compute and network usage of an application, we proposed a new, declarative abstraction called app slice. It allows joint consideration of

compute and network requirements of 5G and edge computing applications. The declarative semantics of the new abstraction are realized by a new app slice runtime, which ensures that the application automatically optimizes the compute and network resource usage on different edge computing environments, and different private or carrier 5G networks. We also implemented a real-world, real-time video analytics application using the app slice abstraction, and performed extensive experiments on a private 5G testbed to demonstrate the positive impact of the proposed abstraction and its runtime system on the performance and resource utilization of applications.

REFERENCES

- [1] K. Cao, Y. Liu, G. Meng, and Q. Sun, "An Overview on Edge Computing Research," *IEEE Access*, vol. 8, pp. 85 714–85 728, 2020. [Online]. Available: <https://ieeexplore.ieee.org/document/9083958/>
- [2] M. H. C. Garcia, A. Molina-Galan, M. Boban, J. Gozalvez, B. Coll-Perales, T. Şahin, and A. Kousaridas, "A tutorial on 5G nr v2x communications," *IEEE Communications Surveys Tutorials*, pp. 1–1, 2021.
- [3] A. Ksentini and P. A. Frangoudis, "Toward slicing-enabled multi-access edge computing in 5g," *IEEE Network*, vol. 34, no. 2, pp. 99–105, 2020.
- [4] "Transformational performance with 5G and edge computing." [Online]. Available: <https://d1.awsstatic.com/Wavelength2020/AWS-5G-edge-Infographic-FINAL-Aug2020-2.pdf>
- [5] N. Hassan, K.-L. Yau, and C. Wu, "Edge computing in 5G: A review," *IEEE Access*, vol. PP, pp. 1–1, 08 2019.
- [6] N. Huin, P. Medagliani, S. Martin, J. Leguay, L. Shi, S. Cai, J. Xu, and H. Shi, "Hard-isolation for network slicing," in *IEEE INFOCOM 2019 - IEEE Conf. on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019, pp. 955–956.
- [7] "Demystifying network slicing." [Online]. Available: <https://ribboncommunications.com/sites/default/files/2020-05/Demystifying-Network-Slicing-WP.pdf>
- [8] S. Wijethilaka and M. Liyanage, "Survey on network slicing for Internet of Things realization in 5G networks," *IEEE Communications Surveys Tutorials*, pp. 1–1, 2021.
- [9] A. Banchs, G. de Veciana, V. Sciancalepore, and X. Costa-Perez, "Resource allocation for network slicing in mobile networks," *IEEE Access*, vol. 8, pp. 214 696–214 706, 2020.
- [10] "Kubernetes." [Online]. Available: <https://kubernetes.io/>
- [11] "Celona." [Online]. Available: <https://celona.io/>
- [12] "iPerf." [Online]. Available: <https://iperf.fr/iperf-doc.php>
- [13] "Multitech." [Online]. Available: <https://www.multitech.com/brands/multiconnect-eccell>
- [14] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative mobile edge computing in 5G networks: New paradigms, scenarios, and challenges," *IEEE Commⁿs Magazine*, vol. 55, no. 4, pp. 54–61, 2017.
- [15] J. Wehrmann, "Movie genre classification: A multi-label approach based on convolutions through time," *Applied Soft Computing*, vol. 61, 2017.
- [16] H. A. Ismail and M. Riasetiawan, "Cpu and memory performance analysis on dynamic and dedicated resource allocation using xenserver in data center environment," in *2016 2nd International Conf. on Science and Technology-Computer (ICST)*, 2016, pp. 17–22.
- [17] "Topology and orchestration specification for cloud applications version 1.0," November, 2013, OASIS Standard. [Online]. Available: <http://docs.oasis-open.org/tosca/TOSCA/v1.0/TOSCA-v1.0.html>
- [18] R. Liu, X. Hai, S. Du, L. Zeng, J. Bai, and J. Liu, "Application of 5G network slicing technology in smart grid," in *2021 IEEE 2nd International Conf. on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 2021, pp. 740–743.
- [19] W. Li, R. Liu, Y. Dai, D. Wang, H. Cai, J. Fan, and Y. Li, "Research on network slicing for smart grid," in *2020 IEEE 10th International Conf. on Electronics Information and Emergency Communication (ICEIEC)*, 2020, pp. 107–110.
- [20] A. Vergutz, G. Noubir, and M. Nogueira, "Reliability for smart healthcare: A network slicing perspective," *IEEE Network*, vol. 34, pp. 91–97, 07 2020.
- [21] Q. Li, G. Wu, A. Papathanassiou, and U. Mukherjee, "An end-to-end network slicing framework for 5G wireless communication systems," 2016. [Online]. Available: <https://arxiv.org/pdf/1608.00572.pdf>