Final Report

NSF Workshop on
Future Directions in Edge Networking

Hawaii, USA
April 20, 2018
Co-located with IEEE INFOCOM

Workshop Chairs
Shigang Chen, University of Florida, sgchen@ufl.edu
Leandros Tassiulas, Yale University, leandros.tassiulas@yale.edu
Joerg Widmer, IMDEA Networks, joerg.widmer@imdea.org

Summary

Edge networking enabled by high performance processing and storage capabilities available at light weight devices at the network edge, and at the same time made necessary by the proliferation of processing intensive time sensitive services that are based on data often collected at the network edge. While research on edge networking is progressing at a rapid pace, many of the concepts, problems, and solution remain to be developed. The purpose of this workshop on edge networking was to bring together researchers from the networking and wireless communities to shed light on important open problems and promising future research directions. The workshop is part of a series initiated by the National Science Foundation and was co-located with IEEE INFOCOM 2018 held at Honolulu, HI, in April 2018. Around 36 attendees participated in the discussion that was organized under the following four thrusts: (1) edge networking in 5G and software defined mobile networks, (2) Internet of things and device-to-device communication, (3) virtualization at network edge and end-to-end, and (4) big data analytics: at edge or back to cloud.

The main discussion points and outcomes of the workshop are:

- For 5G wireless networks, key concepts such as layering in 5G, software defined networking (SDN) and network virtualization, and time granularity need to take edge processing into account. Wireless edge networks should also support long tail services that are customized for a few users. Among others, security and privacy, and access control to local data at the edge are extremely important research topics requiring further investigation.
- The virtualization of services requires cross-layer optimization solutions addressing networking, computing and storage, as well as security and privacy solutions for edge clouds. At the same time, virtualization of access networks and massive IoT infrastructure at the edge are key enablers for a variety of new applications.
- The Internet of Things is a key driver of edge networking, requiring new architectures for edge computing, incentive mechanisms to encourage D2D, and other new enabling ideas such as blockchain and consensus-building tools. Again, privacy and security are extremely in this context, along with resource modeling and data representation/reduction with machine learning based processing. Open testbeds and data sets that provide rich physical layer information will be particularly valuable.
- Finally, AI applications and services will be the main driver of edge computing bringing challenges such as partitioning analytics functionality between edge

computing and cloud computing, analyzing distributed data sources across networks, security and privacy issues of sharing or analyzing data across multiple devices, architectures that promote private edge-computing providers, and mechanisms to share data analytics resources between different edge providers.

The above is a sample of an abundance of novel research problems generated at the edge that are expected to challenge the networking community the years to come.

1. Introduction

Computing and networking have been evolving together since the dawn of computers. Before the emergence of digital networks, centralized mainframe computers dominated, with time-sharing OS supporting multiple user terminals. The advent of PCs and local area networks gave rise to the concept of distributed computing, which proliferated with the development of the Internet and the client-server model. As world-wide web, e-commerce, digital entertainment and countless applications drew virtually everyone into the cyber world, the economics of scale played its role during the past decade in centralizing the computing hardware, software, infrastructure and IT support to huge data centers in the form of cloud computing. In recent years, driven by the need of moving computation and communication closer to the location of applications, this cyclic trend between centralized computing and decentralized computing starts to point back from datacenters towards the edge of the Internet. This form of computing is often referred to as edge computing, or maybe called edge networking if the focus is tilted towards the networking aspect.

Edge networking is in its early stage of research, with its basic concepts, problem scope, theoretical foundation, solution paths, and experimental tools to be defined or developed. The purpose of this workshop is to bring together researchers from the networking and wireless communities who are interested in this subject to shed light on future directions in edge networking through panels and topic discussions. The workshop was the second one in a series that was initiated by Dr. Thyaga Nandagopal and Dr. Wenjing Lou from the National Science Foundation in 2017. It was co-located with IEEE INFOCOM 2018 held at Honolulu, HI, in April 2018.

To put it in context, a related NSF-sponsored workshop on Grand Challenges in Edge Computing was held at Washington DC in October 2016, co-chaired Dr. Mung Chiang and Dr. Weisong Shi. The workshop took a “vertical” approach to identify top five grand challenges in each of the three areas: Applications, Architecture, and Capabilities and Services. The workshop summary can be found at http://iot.eng.wayne.edu/edge/NSF%20Edge%20Workshop%20Report.pdf

Our workshop took a “horizontal” approach to look at the connection between edge networking and other existing research areas. This approach is due to our belief that edge networking does not come in a vacuum. It is a new computing/networking paradigm addressing application needs that cannot be adequately met by the existing system models. It offers great research opportunities that transform existing models by embracing the basic concept of generating data at the Internet edge and moving computing/communication towards the edge. The one-day workshop hosted two panels, each covering two topics, and four breakout thrust discussions, each for one topic. The topics include (1) edge networking in 5G and software defined mobile networks, (2) Internet of things and device-to-device communication, (3) virtualization at network edge and end-to-end, and (4) big data analytics: at edge or back to cloud. In this report, for each thrust discussion, we begin with its goal which is followed by the summary of the breakout discussion and then followed the action plan recommended by the participants.

2. Thrust Discussions

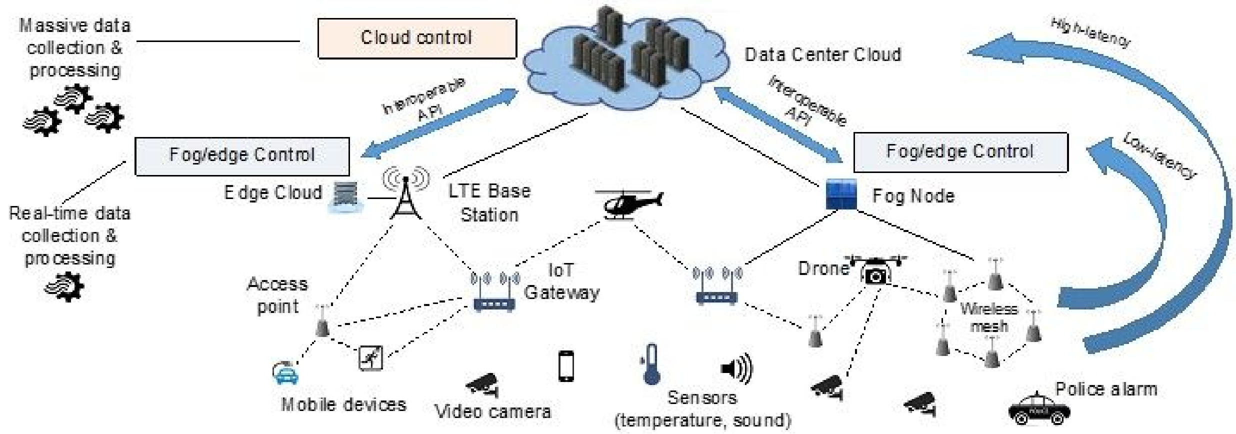
We present the details of the discussions from the main presentations, breakout sessions, and panels on edge networking in the context of 5G, SDN, IoT, D2D, virtualization and big data analytics.

2.1 5G and Beyond & Software Defined Mobile Networks

Lead: Cedric Westphal, Scribe: Jiasi Chen

Goal: 5G wireless communication systems encompass all current innovations in wireless including key developments in networking like Software Defined Networking and Network Function Virtualization (SDN/NFV). The role of SDN/NFV in 5G developments in particular were the focus of discussions. This thrust posed the following questions:

- Is 5G ready for edge computing?
- What is the current status of SDN for mobile networks? Limitations and advantages?
- What functions to bring to the edge? What are the API’s? Slicing/NFV/Caching
- What are the use cases and the business model that will drive the transition to the edge
- As user functions move to the edge how the core network will change and the role of internet will be redefined?



Summary of discussions: There is a need to fully understand how software-defined networks can play a role in mobile networks, as well as the role that mobile networks can play as key members of the network edge.

1. Adapting SDN for mobile networks: SDN is built on the premise of separating the control and data plane, but its use and value in mobile networks requires careful consideration. Wireless networks have different characteristics than the wired domain where SDN is typically applied; for example, full isolation is not possible in wireless scenarios, the time granularity of decisions is finer, and variability is exploited to maximize performance. Can key abstractions of SDN, such as match/action, be applied to 5G networks? One possibility is to map match/action in the wired domain to assigning flows to bearers in the wireless domain. Other, wireless-specific fields may need to be defined for a 5G SDN flow, such as link status and broadcast, or even defining new header fields to enable clients to connect to multiple radio access technologies.

Should SDN be used in the air interface and/or the 5G core network? The appeal of SDN lies in its ability to define new interfaces and interoperate with the network. Wherever it is used, SDN must co-exist with other edge protocols, such as ad hoc protocols. The 5G standard is moving towards separation of the control and the user plane and thus SDN may mesh well. However, SDN is not tied to a specific wireless technology and may be useful in mobile networks beyond 5G.

Virtualization and network slicing will be key components of a SDN-enabled mobile network. Network slicing will enable flows to be granted different classes of service, such as ultra-low latency and ultra-reliable service. Which layer of the network stack should these slices start? In an end-to-end approach, the slice starts at the client and going through the network to the edge/cloud data center, and back. Alternatively, the slice may start at the network layer. Network slicing has some historical similarity to

VPNs, which were defined for the core network with a focus on bandwidth; now, network slices should be defined for the wireless network and encompass bandwidth, computation, and storage.

Non-commercial 5G may be used in certain applications, which may result in different design choices for SDN. For example, some applications with backhaul connections (e.g., Navy sailors browsing the Internet on multiple ships) naturally have a central access point and thus a centralized approach may be appropriate; whereas other applications (e.g., soldiers on a battlefield) exist in a micro-cloud environment without a backhaul connection. Developing SDN for infrastructure-less data transfer, with heterogeneous devices and environments, is an interesting area of exploration. Minimizing exfiltration of information to the cloud allows communities to thrive using basic edge computing (local) services without necessarily relying on the larger ecosystem of services that depend on the cloud - as the edge compute infrastructure could end up being lower in cost and complexity.

2. Defining the right abstractions: A recurring theme in the discussion was the need to design the right abstractions for SDN. The proper abstractions will help the edge infrastructure support applications, just as the cloud infrastructure does today, as well as enable easy orchestration of resources. What abstractions, APIs, and hierarchies are needed? How can we help translate requests from system managers (e.g., “add bandwidth to this slice” or “improve QoE”) into an actionable policy using an intent manager? There may be multiple policies available, such as scheduling, beamforming, and coding schemes. How to choose the right policy? Data analytics may help here. Platforms such as PAWR can be used to not only experiment with lower-layer technical issues but also with policy implementation.

There are still many challenges in the physical layer, so it may be difficult to determine the right abstractions when the physical layer is still evolving. Defining the right abstractions will also require help from computer scientists to determine the appropriate layers of the network stack to act on, and what APIs are needed to support the edge infrastructure (e.g., to dynamically load edge applications).

3. Data plane models: In order for a 5G SDN to orchestrate and easily manipulate resources, an appropriate model of the data plane is needed. In the current SDN control plane, manipulation of the data plane happens at the flow level, but further time granularity (e.g., at the sub-millisecond level for scheduling) may be needed if SDN is used at the air interface. Data plane models include theories of fairness and quality-of-service, which should apply to single users as well as groups of users with particular service-level agreements, and thus require different network management. Business and economic models of the data plane may also be needed in order to support the overall edge ecosystem.

Action plan: Based on the discussion, the following action items are recommended:

- Research is needed on the design of software-defined mobile networks, including key questions of layering, network virtualization, centralization, and time granularity.
- The right abstractions, APIs, protocol stack, and policy selection mechanisms need be developed in order for wireless edge networks to be successful.
- The edge need to be able to support “long tail” services, that is to be able to instantiate customized services for a few users; this requires new methods to cheaply and efficiently deploy services, along with new business models to support these services
- Wireless resources are still precious; therefore, well-defined data models are needed for the SDN control plane to be able to easily manipulate and orchestrate data.
- Key areas of investigation are the benefit of edge computing for security and privacy, and how to control access and leverage the data locally.

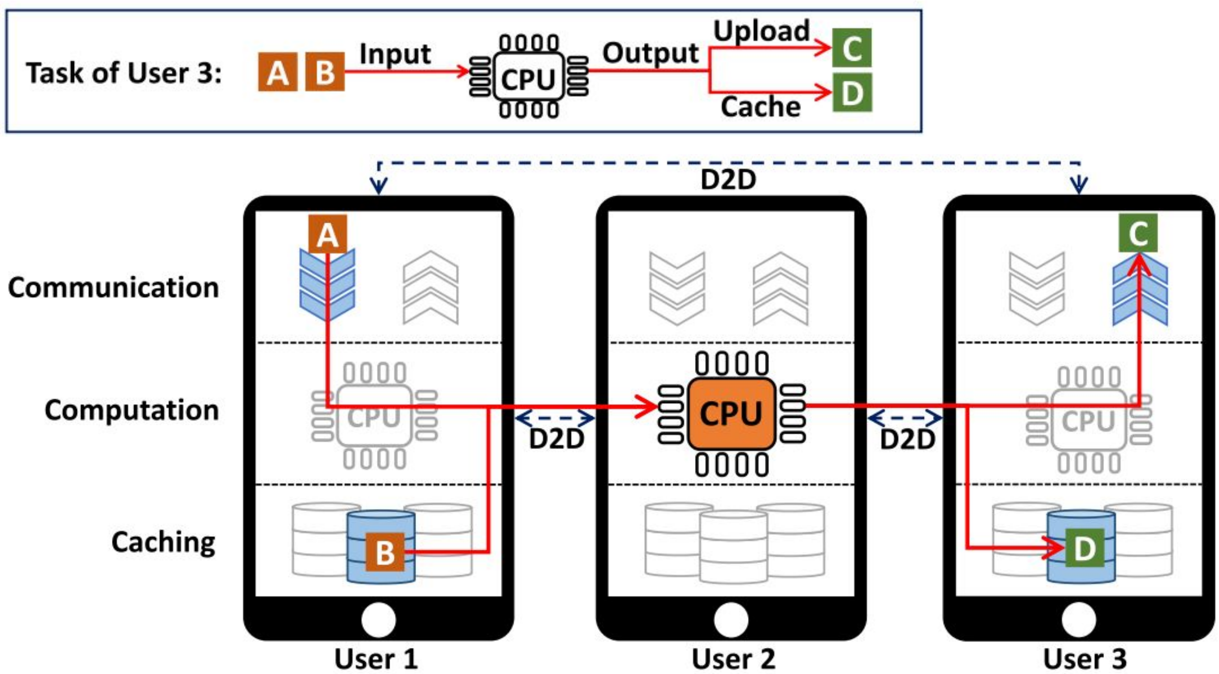
2.2 Internet of Things and Device-to-Device Communications

Lead: Jianwei Huang (Chinese University of Hong Kong)
Scribe: Yan Wang (Binghamton University)

Goal: The Internet of Things (IoT) and device-to-device (D2D) communications provide key support for future edge computing. D2D enables nearby mobile devices in an IoT system to share their resources for cooperative task execution. Such resource sharing can effectively pool the mobile users’ heterogeneous resources and improve the overall network performance. This thrust posed the following questions:

- What are the possible new architectures for D2D communications and resource sharing, especially in the context of IoT applications for edge computing?
- What are the new methodologies to solve technological and incentive issues?
- What are the privacy and security concerns in D2D communications and edge computing?
- How to evaluate the performance of D2D communications?
- What are the system-level challenges in implementing and evaluating D2D communications and IoT?

Summary of Discussion: The emerging paradigm of Internet-of-Things and the growing computing capability of commodity mobile devices have encouraged people to use mobile devices to tackle various complicated tasks, such as online gaming, data processing, and augmented reality, at the Internet edge. These tasks may require significant communication resources (for data downloading and uploading), computation resources (for data processing), and caching resources (for data storage and retrieval), named as 3C resources. One challenge is due to heterogeneity in both resources and service requirements. To resolve the problem of resource/service mismatch, one possible solution is to allow nearby mobile devices to share their resources for cooperative task executions through local device-to-device (D2D) connections (e.g., Bluetooth and WiFi Direct).



3C Resource Sharing based on D2D Cooperation (image from Ming Tang, Lin Gao, and Jianwei Huang, "Enabling Edge Cooperation in Tactile Internet via 3C Resource Sharing," IEEE Journal on Selected Area in Communications, Special issue on Emerging Technologies in Tactile Internet and Backhaul/Fronthaul networks 2018)

Architecture and Ecosystem: Regarding the future architectures for D2D communications and resource sharing in IoT systems, some participants thought that fog computing can be a suitable general framework that connects cloud computing, edge computing, and everything in between. Much research is needed to define the boundaries of these technologies, which remain vague today. One of the challenges in edge-computing architecture design is identifying and sharing the 3C resources (communications, computation and caching resources). To achieve this, one approach is shifting from “task-centric” to “resource-centric”. For example, a task can be decomposed into subtasks involving different resources, which can be completed at different locations of the network. An efficient system for resource naming and localization is critical under such a design.

Also importantly, the research community needs to build an ecosystem for collaborative work in edge computing. Such an ecosystem defines the naming and ownership of resources, provides the mechanisms to promote resource sharing, and establishes the business models for diverse application scenarios. Moreover, we should think about how to properly incentivize different stakeholders to collaborate and develop the enabling technologies. For example, it may be possible to use blockchain to facilitate distributed coordination, consensus formation, and micropayment transactions among participants. Auction and contract mechanisms can be used to tackle situations with incomplete network/channel information. These incentive mechanisms need to be designed and optimized under various tradeoffs in complexity, robustness, security and performance. As a first step, there is a critical need for an open platform that enables the trials of different research and business models.

Resource and Performance: A major technical challenge is how to enable a large number of IoT devices to effectively share wireless resources including the emerging 5G. The 5G communication networks promise to support massively broadened medium access, but latency can become a problem when a base station is far away from its mobile nodes or is experiencing a high load. One solution is to offload some communication and computing needs to nearby edge nodes, which will require coordination between edge computing and 5G technologies, especially when the users are on the move.

It is important to develop performance metrics for evaluating the effectiveness of D2D and edge computing. Some participants believed that proper modeling of the traffic and corresponding latency requirements will greatly facilitate the assessment to the quality of D2D communications. Data reduction and reformulation involving machine learning based processing are also valuable because IoT and edge computing will likely deal with significant data in networks with limited bandwidth. In addition, energy consumption in IoT and sensing networks is a challenging problem for research.

For resource discovery, the discussion pointed out the need for a good understanding on modeling various resources in the networks. Resources available in D2D and edge computing are not be limited to traditional computer hardware such as processing units, memory and bandwidth. Other metrics such as sensing modalities and data quality should also be considered in the classification of resources. Naming and identification of resources serve a critical role in resource modeling in the context of edge computing.

Regarding system implementation and testbeds for edge computing based on D2D and IoT, some participants noticed that the accessibility of various protocol stacks in current mobile OS is very limited. It is nearly impossible to implement edge computing protocols directly on commodity mobile devices. To meet such challenges in the system implementation, the research community is in urgent need of open testbeds that provide flexible access to physical layer data and system level functions.

Security and Privacy : For resource and information sharing in IoT and D2D communications, the privacy and security of user data become crtitical issues in practice. There is a pressing need for the research community to agree on proper privacy/security models in edge computing. Based on such models, privacy filters may be adopted at edge nodes to control the dissemination of sensitive user data and only allow the flow of information that is necessary to facilitate the edge-computing applications. In addition to data privacy, resource discovery and sharing also have their own unique security concerns. For example, when sharing their resources with nearby devices, mobile nodes may expose the patterns of their resource availability and reveal their local computing activities and user behaviors.

Another interesting issue on privacy and security is related to sensing applications. For IoT systems built on a variety of sensing applications, it may be too costly for each application to maintain its own sensing substructure. If a generic sensing structure is available as a system function, the middle layer can serve as a bridge to connect the application needs with raw data provided from the system layer. There are many underlying security issues to be addressed in such an architecture --- for example, how to deploy access control in the middle layer that allows mobile nodes without direct Internet connections to access the data, and how to support crowd-sourcing while preventing the large-scale temporal-spatial data from being abused for unintended purposes.

Action Plan: Based on the panel presentation and the breakout discussions, the following actions are recommended by the discussion group.

- Both the NSF and the research community should make a collective effort to establish a sustainable ecosystem for D2D and IoT based edge computing.

- Active research is needed on developing new architectures for edge computing, including resource discovery, distinction of edge, core and fog resources, incentive mechanisms for encouraging D2D collaborations, and platforms that support experimentation on new enabling ideas (such as blockchain as an incentive and consensus-building tool).
- Research on privacy and security is extremely important for D2D and IoT. This includes but is not limited to how to design privacy filters on edge nodes to control the flow of sensitive data, how to model data privacy in edge computing, how to balance the capability of technologies and the need for regulations, and how to support security in resource discovery and sharing.
- Resource modeling and data representation/reduction (with machine learning based processing) are important research directions. In addition, latency, traffic measurement, and energy management are also valuable research topics in the context of edge computing.
- Due to the difficulty in accessing the physical layer information from today's commercial mobile OSes, open testbeds and data sets that provide rich physical layer information are urgently needed in the community.

2.3 Virtualization at Network Edge and End-to-end

Lead: Ulas C. Kozat; Scribe: Konstantinos Poularakis

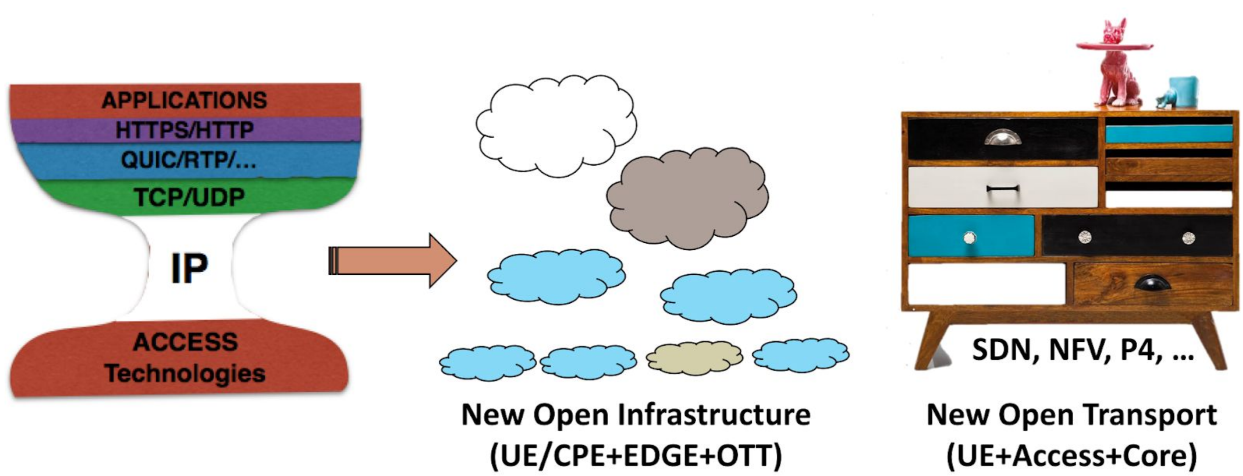
Goal: Virtualization is a broad topic that can have quite different usage models and optimization criteria. As virtualization is pushed towards the edge, it transforms the whole Information and Communication Technology (ICT) infrastructure and how we deploy applications, features, and services. This thrust posed the following questions:

- When we push virtualization to the network edge? How do the existing research problems differ? Do they become more challenging to be addressed?
- What are the new challenges that emerge due to pushing virtualization to the network edge?
- What is the edge? Where do we start/stop virtualization?
- What do we virtualize, for whom do we virtualize, how do we virtualize?
- How does the principle of software-defined networking come into play when virtualization is pushed to the network edge, as compared to when it is deployed in large centralized data centers?
- What type of new applications and usage models emerge as we push virtualization to the network edge?

- Do we have the right set of theories to address the design challenges and trade-offs for network/infrastructure virtualization?

Summary of discussions: As we move forward into 5G era, the boundaries between communication, computation, and storage will get more blurred. Creating a seamless infrastructure for applications starting from the edge devices (e.g., end user equipment, customer premise equipment, vehicles, robots, etc.) up to the centralized data centers will take advantage of cloudified access and core networks as well as a transport fabric built upon software defined networking (SDN) principles. On one side of the token, service providers can offload the computation and communication to edge clouds that require high bandwidth, low latency or highly reliable communications with edge devices. On the other side of the token, edge devices can be provided a seemingly “infinite” pool of virtual memory, virtual computation cores, and virtual transport resources that extends from the local device to the edge cloud and beyond. Virtualization frees applications from the physical constraints of the underlying infrastructure and simplifies deployment across the hard boundaries of different clouds.

When virtualization is coupled with network slicing and edge cloud, the “thin waist” model of the Internet is transformed into a new model where it becomes possible to customize the end to end network stack to optimize the application performance under the existing (potentially dynamically changing) resource constraints and workload characteristics. Particularly delivering high bandwidth, low latency or highly reliable communications is expected to require a deeper control and coordination between the end points and the network functions. SDN and network programmability will be indispensable for such a paradigm.



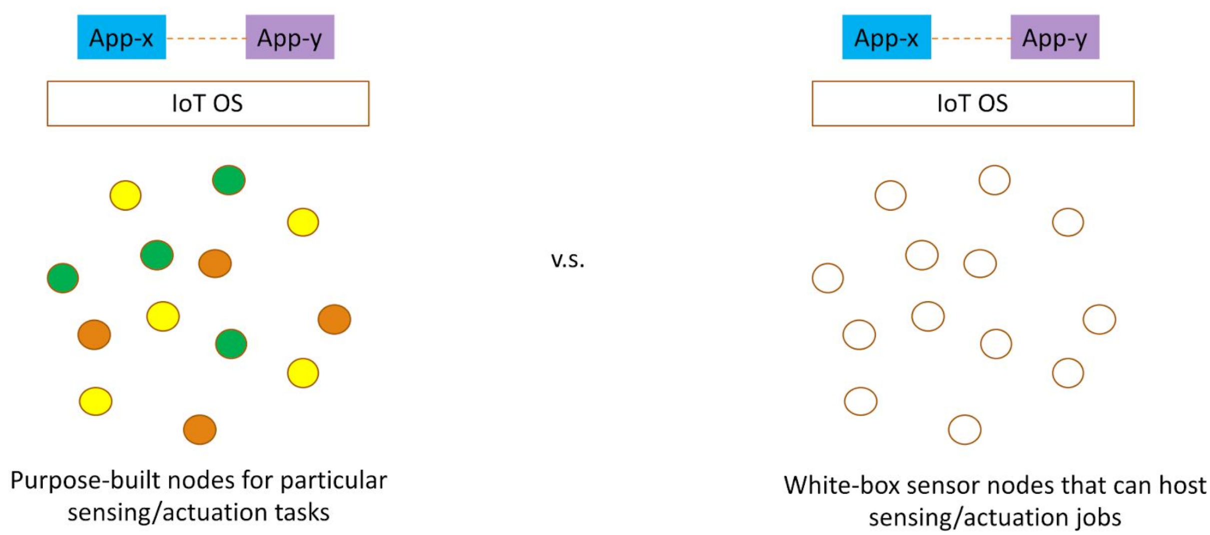
From Narrow Waist to New Cloud & Open Chest

The definition of edge and use cases change the scope of virtualization. In data centers, virtualization is used to address several pain points including (but not limited to): (1) running legacy software and networking stack unmodified over the continuously changing software and networking stack in the data centers; (2) decoupling applications and services from the hardware or physical resource constraints; (3) package all software dependencies in one image; (4) isolate workloads (or tenants who generate workloads) and software from each other; (5) utilize commoditized processors for computing. Moving virtualization to the edge however means running more real-time applications and network functions that may be sensitive to the overheads of virtualization that provides all the initial benefits. To increase performance, system architects and researchers start (i) relying on particular hardware that can natively support virtualization, provides acceleration or offloading; (ii) relaxing isolation features (e.g., use the same OS kernel); or (iii) cross-layer optimize the software and networking stack throwing away all the unwanted features to support a single-purpose application stack. The result is that not only applications are tied to a particular hardware configuration but also to a particular stack of software development, configuration and management system. We are missing a principled approach and a theory of virtualization that address the trade-offs among security, isolation, universality, performance and scalability.

Compared to traditional clouds, edge clouds can improve latency and privacy (of data and computation). Edge clouds individually provide a limited resource footprint, but collectively they can rival centralized clouds. How to achieve these collective gains however poses a big challenge. Also, providing a trusted computing environment over many distributed (possibly loosely federated) edge clouds remains as another major issue. When users are mobile, edge computing and virtualization need to support mobility not as an aftermath, but as a first class requirement. Edge virtualization is expected to bring new security threats that can bring down the edge clouds and access networks. For instance, when third party virtualized network functions are deployed, new vulnerabilities for network security and data privacy surface.

Virtualization of the mobile/wireless access networks opens up new questions about what it means to virtualize wireless transport related resources such as spectrum, antennas, baseband processing, and even power resources. Different than multi-tenancy that already exists today, virtualized resources gives the impression to the users that they own their dedicated resources and they can decide how to use use/control these resources. Virtualization can be adopted at a given layer or can be applied in a nested fashion to cover different operational needs. Which way of wireless access virtualization is better than the others requires a common evaluation and testbed environment that is capable of supporting many virtualization methods over the same testbed.

Virtualization of IoT networks is another avenue fertile for virtualizing sensing and actuation jobs. Central clouds have created several deployments for IoT application developers to pull and process data from IoT devices. A new trend is occurring towards opening devices at homes, public spaces, offices and factory floors to access locally collected data and provide value add services. Via virtualizing the sensing and actuation capacities of these IoT devices together with the communication network among themselves or between them and the nearby edge clouds, more intelligent homes, factories, parks, cities, etc., can be developed by many businesses, communities, and individuals.



An Edge Case, Wireless IoT

Another interesting point raised during the discussion is that whether we can virtualize a person. In the simplest form, this can be a personal assistant running in the cloud. This notion has at one level similar to the current personal assistants supported by cloud providers that serve many users and build per-user models for personalized experience and recommendations. When pushed to the limit, however, having a conscious self that runs and roams around in the cloud ether that can mirror a human being in her/his absence is far beyond from what we have today commercially or experimentally.

Network neutrality has been also discussed in the context of access providers. NFV and SDN challenge the transport centric notion of network neutrality. How to democratize the computing and storage capacity of edge clouds while protecting the interests of users/consumers is posed as a new problem. Extending fairness notions to cover a larger set of resources in the edge cloud is needed. Modeling the inter dependencies across resources such as memory, CPU, link bandwidth, etc., to provide a network service level agreement as well as modeling the performance dependencies on the particular stack of virtualized network infrastructure (e.g., processor types, hardware/software accelerators, hypervisors, operating system, etc.) become critical to optimize the network performance and efficiency.

Action Plan: Based on the discussions in the panel and breakout session, the group proposes the following action plan:

- As networking becomes very much intertwined with computing and storage, research community must redefine metrics for optimization, revisit fairness, redefine scaling laws, and work on new set of cross-layer optimization problems.
- More focused research is needed to develop security and privacy threat models along with solutions for edge clouds and virtualization.
- With virtualization and programmability, networks are no longer fixed pipes and how to change the pipe behavior together with the end to end techniques (e.g., rate control, source coding, multi-path, etc.) opens up new opportunities.
- Virtualization of access networks and massive IoT infrastructure at the edge are critical components for the future of our society from health care management to Industry 4.0 to autonomous driving and smart cities. More efforts are needed from the research community to make this shift happen faster.
- Experimentation tools and testbeds are essential components to operationalize research ideas on edge virtualization. The existing efforts must be sustained and grown further.

2.4 Big Data Analytics: at Edge or Back to Cloud?

Lead: Anwar Walid (Nokia Bell Labs)
Scribe: Carlee Joe-Wong (Carnegie Mellon Univ.)

Goal: Traditional forms of big data analytics do not consider distributed sources of data or constraints on the analysis algorithms, e.g., a need to deliver results quickly. Instead, they are mainly concerned with ensuring that the results are as accurate as possible. In many network applications, however, data are often generated from multiple sources, such as different points along a flow's path, and may need to be analyzed quickly in order to deliver actionable results. Also, due to shear volume or privacy concerns, it may not be feasible to gather data in a central location for processing. Conversely, many big data applications with multiple data sources and delay constraints will need networking support to connect the devices and ensure that they can communicate with each other timely. This discussion focused on determining the important research questions to answer in this space and identifying both challenges and opportunities for networking and big data analytics.

Summary of Discussion: The discussion tried to explore the intersection between big data analytics and edge computing by centering around two themes: using distributed data analytics as a means to solve networking problems and developing networking-based solutions for distributed data analytics. In addition, there was significant discussion about the types of research questions that the networking community should address. This summary is structured around four major topics: (1) big data analytics for networking, (2) networking and edge computing for distributed data analytics, (3) challenges in distributed analytics and learning, and (4) standardized datasets and platforms. Common concerns across these questions were security and privacy, developing application-specific or generic solutions, and the need for platforms and datasets that are open to everyone in the community, instead of siloed solutions.

Data Analytics for Networking Problems: While many networking problems can be solved with off-the-shelf data analytics methods, e.g., traffic prediction and dimensionality reduction of large network datasets, other problems raise new challenges that can benefit from edge computing due to their distributed nature, where the concept of edge is generalized to include both end systems and network devices, with respect to centralized cloud-based data analytics. Network diagnosis, for example, requires analyzing data from multiple points along a flow's path through a network. Other network problems involve both inferring insights from distributed data and actuating devices based on these inferences. Many multi-dimensional problems cannot be solved with model-based approaches. Discretization may help to reduce the complexity, and prior work has shown that machine learning and reinforcement learning are applicable to network problems such as finding optimal schedulers, with promising results that match well with traditional approaches. Indeed, learning-based approaches in general capture more detail than model-based approaches, which tend to use simplified models of reality. Yet this introduces a new challenge of understanding the semantics of data throughout a network. While cross-layer optimization has been studied for many years, there is still no common language to represent features and variables across network layers.

Networking for Distributed Data Analytics: Distributed analytics will require networks to connect the devices across which data are distributed and analytics algorithms are executed. This will require the participation of private edge resource providers who own these devices. Some discussion participants pointed out that there might be an incentive for national cloud providers to become edge providers, which allows the opportunity of integrating cloud computing and edge computing in distributed data analytics and in addition stimulates competition between providers. Multiple edge providers for distributed data analytics naturally raise the question of pricing their resources. For instance, different components of edge analytics may contribute different levels of accuracy to the distributed analytics algorithms, which may induce different prices for different types of resources.

Privacy and security are important research topics for distributed data analytics. Secure analytics such as multi-party computation must balance between performance and security by making tradeoff among a variety of factors including computation overhead, communication latency, and potential security pitfalls. Data provenance issues can also arise in distributed analytics where one must determine who owns or has the right to view different pieces of data. Care must be taken to prevent data aggregators from reverse-engineering some data that originates from other sources, as in differential privacy. Distributed data analytics also involve authentication and authorization mechanisms, as well as trust management among devices.

One model of networking for distributed analytics draws analogy from WiFi service provisioning, where different organizations operate their own WiFi access points and service providers lease this existing infrastructure instead of building their own. While edge computing is significantly more complicated than WiFi access, this analogy may lead to some starting points to solve access problems for edge resources. For instance, many challenges faced by the WiFi service, such as authentication, dynamic access control, mobility and cross-layer optimization, need to be dealt with in edge computing as well. One interesting research problem is to integrate edge computing and cloud computing in distributed data analytics, with different components of learning and analytics algorithms being placed across edge and cloud devices. Applications must decide when to use edge or cloud resources, accounting for performance, privacy and security, whose particular trade-offs will depend on the particular application needs. One idea is to train models in the cloud and port them to the edge for execution.

Challenges in Distributed Analytics and Learning: Distributed computing architectures exist today for data analytics, including Tensorflow and distributed parameter server architectures. These architectures are generally deployed for multiple servers in a compute cluster, letting applications take advantage of multiple CPUs. While these architectures could be naively deployed for edge computing, they do not take into account the much larger transmission delays in edge architectures compared to computing clusters. This difference changes the performance tradeoffs in distributing data analytics. In particular, many parameters need to be exchanged between different servers in edge computing, which may result in significant overhead and delay. Virtual or augmented reality applications present particularly challenging scenarios, as they have stringent latency constraints. Some models have implemented distributed data analytics with local training (e.g., SVM) on Raspberry Pis with global parameter exchange, which provide possible initial solutions to this research challenge.

One common challenge faced by parameter server architectures is asynchronous updates that can cause delays in model training. Many papers look at the resulting latency vs. accuracy tradeoffs, as well as performance and resource utilization tradeoffs that consider CPU, bandwidth, memory, power, latency, etc. In the dge computing context, it may be possible to execute part of a learning algorithm at the end device where data is generated, e.g., by running a pre-learning or filtering algorithm. Fingerprinting is a potential application of this idea. Other ways that existing distributed learning architectures can be modified for edge computing settings include discretizing the weights in the machine learning architectures or developing specialized chips (for facial recognition as an example).

Common Datasets: The main conference at INFOCOM 2018 featured a panel on machine learning and networking. One of the challenges discussed during the panel, which also came up during the breakout discussion of this workshop session, is the need for common datasets on which to evaluate machine learning or data analytics solutions to networking problems. These datasets can help ensure research reproducibility. Some datasets from CAIDA on network traffic are widely used, but these are far from standardized. Another challenge is that the landscape of networking research changes too fast for a common benchmark to remain valid for long. This requires an evolving mechanism to refresh the datasets with forward looking because some interesting networking problems and new algorithms arise from analyzing datasets from emerging networking paradigms.

Action Plan: The discussion group recommended several specific research topics to investigate, based on the panel and breakout discussions:

- New research questions that come from analyzing distributed data sources across networks
- Security and privacy issues that accompany sharing or analyzing data across multiple devices

- Reinforcement learning for networking problems, especially for the Internet of things
- Architectures that promote private edge-computing providers, including incentives for them to participate in analyzing data across different edge devices
- Mechanisms, including pricing, for applications to share data analytics resources between devices from different edge providers
- Partitioning analytics functionality between edge computing and cloud computing
- Role of industry in facilitating big data analytics research, e.g., providing standard datasets and platforms to test research ideas

3. Conclusions

Edge networking is progressing fast as a natural evolution of computing and networking. The computing technology advances on one hand make feasible high performance processing and storage capabilities in light weight devices that may reside at the network edge. The proliferation of data at the network edge on the other hand as well as the proliferation of processing intensive time sensitive services based on this data make edge processing the only option for delivering the services. Networking and computing are the key enablers of that development and co-evolve to achieve this vision. The development of 5G wireless networks are driven to a large extent by this vision. More specifically: a) Key questions of layering in 5G, network virtualization, centralization, and time granularity need to take edge processing into account; b) Wireless edge network requirements will drive the right abstractions, APIs, protocol stack, and policy selection mechanisms; c) The edge need to be able to support "long tail" services, that is to be able to instantiate customized services for a few users; this requires new methods to cheaply and efficiently deploy services, along with new business models to support these services; d) Wireless resources are still precious; therefore, well-defined data models are needed for the SDN control plane to be able to easily manipulate and orchestrate data; e) Key areas of investigation are the benefit of edge computing for security and privacy, and how to control access and leverage the data locally.

Virtualization of services is moving fast both in computing and networking, facilitated by the softwarization of the operations. This trend will accelerate and new challenges need to be addressed: a) Networking, computing and storage become much more intertwined and the metrics for optimization need to be redefined including the notion of fairness, the scaling laws as well as the cross-layer optimization problems in the new setting; b) Security and privacy threat models along with solutions for edge clouds and virtualization need to be developed; c) With the evolution of virtualization and programmability, the networks become much more agile and an important challenge is how to take advantage of that with novel end-to-end techniques (e.g., rate control, source coding, multi-path, etc.); d) Virtualization of access networks and massive IoT infrastructure at the edge are key enablers in a variety of applications including health care management, autonomous driving and smart cities; e) Experimentation tools and testbeds are essential components to operationalize research ideas on edge virtualization.

The Internet of Things is a key driver of edge networking and places specific challenges: a) New architectures for edge computing are necessitated including resource discovery, distinction of edge, core and fog resources, incentive mechanisms for encouraging D2D collaborations, and platforms that support experimentation on new enabling ideas (such as blockchain as an incentive and consensus-building tool); b) Privacy and security is extremely important for D2D and IoT including designing privacy filters on edge nodes to control the flow of sensitive data, modeling data privacy in edge computing, balancing the capability of technologies and the need for regulations, and supporting security in resource discovery and sharing; c) Resource modeling and data representation/reduction (with machine learning based processing) are important research directions; d) Latency, traffic measurement, and energy management are also valuable research topics in the context of edge computing; e) Open testbeds and data sets that provide rich physical layer information will be particularly valuable, due to the difficulty in accessing the physical layer information from today's commercial mobile OSes.

AI applications and services will be the main driver of edge computing. These applications pose their own requirements that translate to research challenges for the network edge: a) Partitioning analytics functionality between edge computing and cloud computing; b) Analyzing distributed data sources across networks; c) Security and privacy issues that accompany sharing or analyzing data across multiple devices; d) Reinforcement learning for networking problems, especially for the Internet of thing; e) Architectures that promote private edge-computing providers, including incentives for them to participate in analyzing data across different edge devices; f) Mechanism design, including pricing, for applications to share data analytics resources between devices from different edge providers.

The above is a sample of an abundance of novel research problems generated at the edge that are expected to challenge the networking community the years to come.

Appendix A. Workshop Program

7:15 – 8:00am Lehua Lounge
Breakfast and Registration

8:00 am – 8:15 am Lehua Suite
Opening Remarks, Thyaga Nandagopal, John Brassil and Sandip Kundu from NSF

8:15am – 8:30 am Lehua Suite
Workshop Overview by Workshop Co-chairs, Shigang Chen, Leandros Tassiulas, Joerg Widmer

8:30 am – 9:30 am Lehua Suite
Panel Discussion on Topics 1 and 2
Yanchao Zhang (chair), Wenye Wang, Cedric Westphal, Guohong Cao
Topic 1: 5G and Beyond & Software Defined Mobile Networks
Topic 2: Internet of Things and Device-to-Device Communication

9:30 am – 9:45am Lehua Lounge
Coffee break

9:45 am – 11:15am
Breakout discussion (Session 1 for Topics 1 and 2)
· Nautilus 1
Topic 1: 5G and Beyond & Software Defined Mobile Networks
Lead: Cedric Westphal, Scribe: Jiasi Chen
· Nautilus 2

Topic 2: Internet of Things and Device-to-Device Communication
Lead: Jianwei Huang, Scribe: Yan Wang

11:15 am – 11:45pm Lehua Suite
Breakout groups reconvene (Topic leads summarize the discussions on Topics 1 and 2)

11:45 am – 12:45 pm Sea Pearl 1-4
Lunch

12:45 pm – 1:45 pm Lehua Suite
Panel Discussion on Topics 3 and 4
Panelists: Baochun Li (chair), Ulas Kozat, Anwar Walid, Wenjing Lou
Topic 3: Virtualization at Network Edge and End-to-End
Topic 4: Big Data Analytics: at Edge or Back to Cloud?

1:45 pm – 2:00 pm Lehua Lounge
Coffee break

2:00 pm – 3:30pm
Breakout discussion (Session 2 for Topics 3 and 4)
· Nautilus 1

Topic 3: Virtualization at Network Edge and End-to-End
Lead: Ulas Kozat, Scribe: Konstantinos Poularakis
· Nautilus 2

Topic 4: Big Data Analytics: at Edge or Back to Cloud?
Lead: Anwar Walid, Scribe: Carlee Joe-Wong

3:30 pm – 4:00 pm Lehua Suite
Breakout groups reconvene (Topic leads summarize the discussions on Topics 3 and 4)

4 pm: Adjourn

Appendix B. Workshop Attendees

Anthony Ephremides, University of Maryland, USA
Anwar Walid, Bell Labs
Baochun Li, University of Toronto
Carlee Joe-Wong, CMU
Cedric Westphal, Huawei
Eytan Modiano, MIT, USA
Galen Sasaki, University of Hawaii
Georgios Paschos , Huawei
Guohong Cao, Penn State University
Hongyi Wu, Old Dominion University
Jack Brassil, NSF
Jianwei Huang, The Chinese University of Hong Kong, Hong Kong
Jiasi Chen, UC Riverside
Joerg Widmer, IMDEA Networks
Kevin S Chan, US Army Research Laboratory, USA
Konstantinos Poularakis, Yale University, USA
Leandros Tassiulas, Yale University
Lixia Zhang, UCLA
Randall A Berry, Northwestern University
Sandip Kundu, NSF
Sastry Kompella, Naval Research Laboratory, USA
Shigang Chen, UF
Shivendra Panwar, NYU
Shiwen Mao, Auburn University
Suman Banerjee, University of Wisconsin-Madison
Thyaga Nandagopal, NSF
Tom Hou, Virginia Tech
Ulas Kozat, Huawei
Wenjing Lou, Virginia Tech
Wenye Wang, North Carolina State University
Yan Wang, Binghamton University
Yanchao Zhang, Arizona State University
Yingying Chen, Rutgers University
Yuanyuan Yang, Stony Brook University
Yunhao Liu, Michigan State University
Zhi-Li Zhang, University of Minnesota