Towards Intelligence on the Edge: Restructuring Computing to Enable the Next Generation of the IoT

Maria Gorlatova

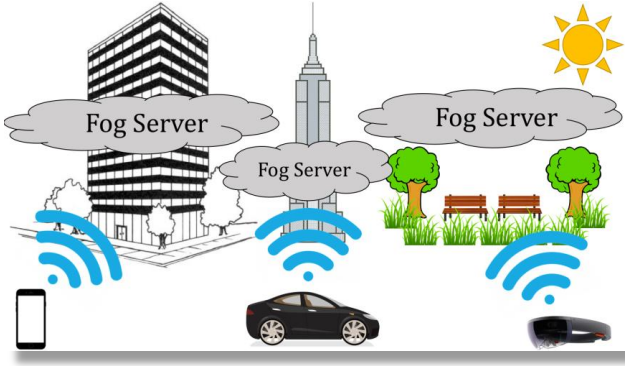
September 25th, 2018

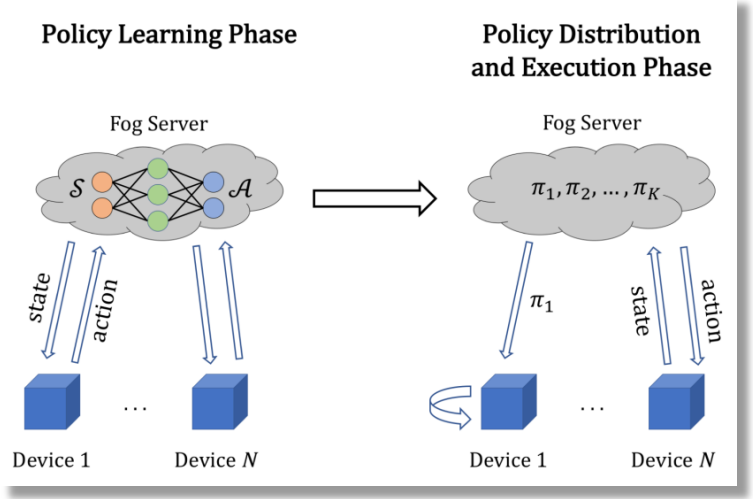
About the Speaker

- Started at Duke University in July 2018
- Previously:
  - Associate Research Scholar, Princeton University, Electrical Engineering
  - Ph.D. Columbia University, Electrical Engineering
  - Industry positions:

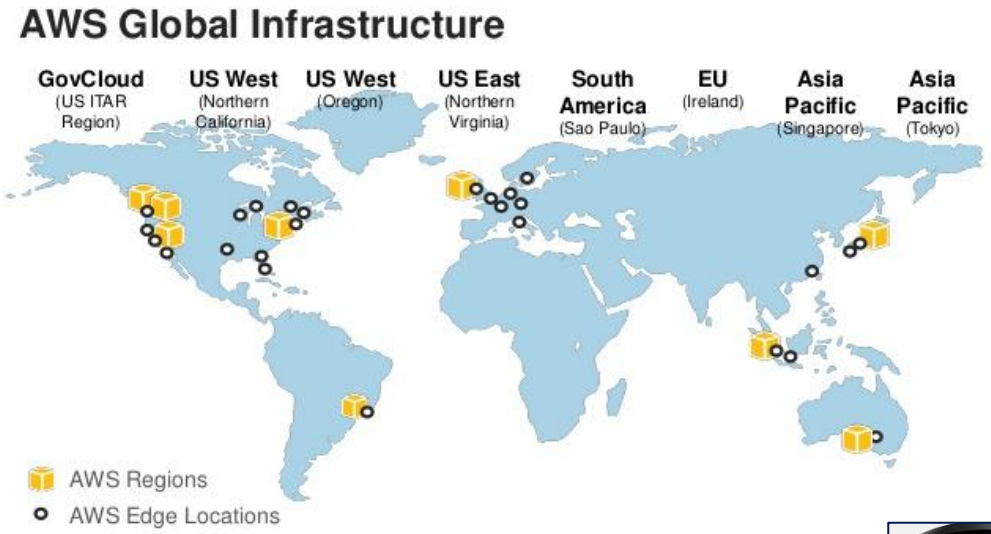
Towards Intelligence on the Edge

- Edge/fog computing
- Characterizing fog substrates
- New capabilities enabled by fog: intelligent **augmented reality**



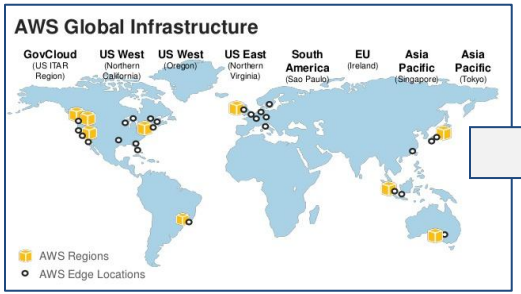


Cloud: Computing in Datacenters

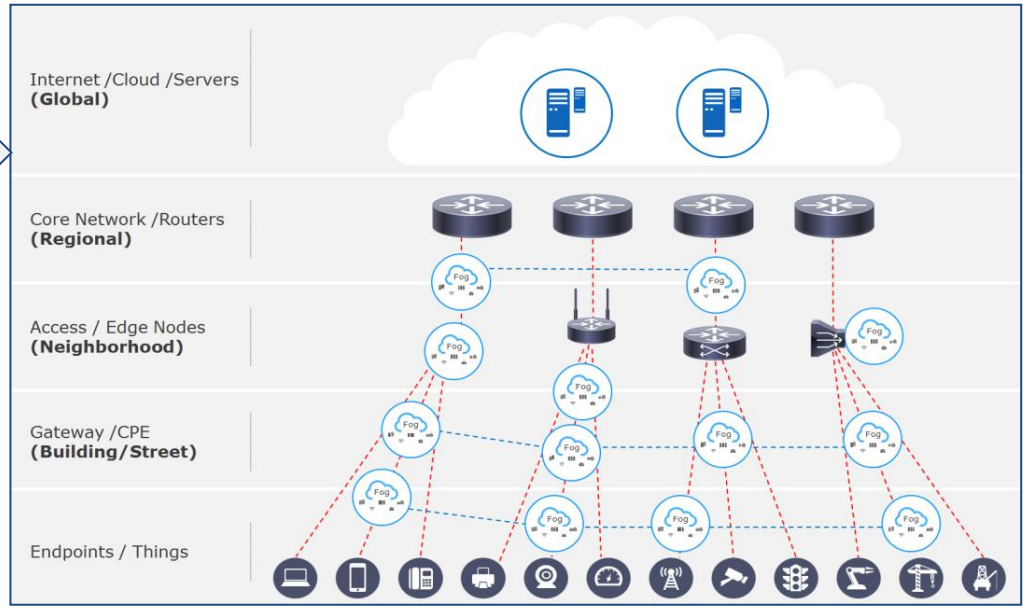


- AWS: 44 locations worldwide, MS Azure: 30
- For emerging applications: **fundamental limitations** in **latency**, **bandwidth**

Edge/Fog: Computing Closer to the Users



- Data processing, business logic, decision-making at multiple points in the hierarchy



Smart city IoT deployments: computing in buildings, neighborhoods, zip codes

Important for Multiple Industries

AWS Greengrass

Cloud IoT Edge ALPHA

Deliver Google AI capabilities at the edge.

REQUEST ACCESS

"Most interesting part of cloud computing"

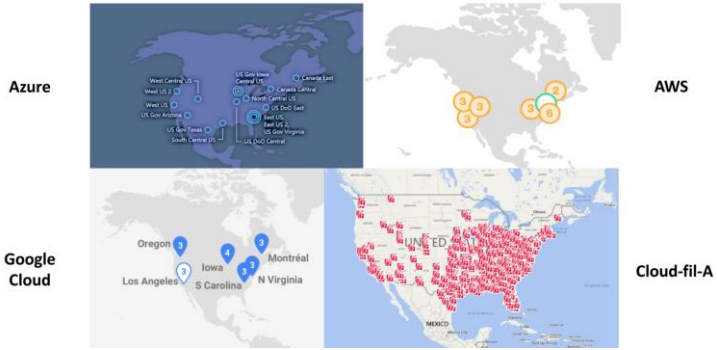
Alibaba, Intel launch joint edge computing platform

Sep. 20, 2018 10:18 AM ET | About: Alibaba Group Holding Limited (BABA) | By: Brandy Betz, SA News Editor

AT&T Foundry Powers Up Edge Computing Test Zone in Silicon Valley to Drive Innovation in 5G Era Feb. 2018

Verizon peels back curtain on edge computing, deep learning for real-time video analytics

by Mike Dano | Mar 29, 2018 12:59pm



New Interdisciplinary Research Area




Hyper-local ⟺ **Local** ⟺ **Regional** ⟺ **Global**

- Differentiating principles: heterogeneity, hierarchy
- Layers with **vastly differing capabilities**
  - Latency, energy, reliability, cost, …
- Layers working **together**
- From embedded systems to cloud management to machine learning

Towards Intelligence on the Edge

PURDUE UNIVERSITY

THE UNIVERSITY OF MELBOURNE

PRINCETON UNIVERSITY





DARPA

amazon

NSF

Microsoft

COMCAST

Towards Intelligence on the Edge

- Characterizing fog
- New capabilities enabled by fog: intelligent augmented reality

- P. Naghizadeh, M. Gorlatova, A. Lan, M. Chiang, On Information Sharing in Multi-Agent Learning, under submission.
- Y. Ruan, L. Zheng, M. Gorlatova, M. Chiang, C. Joe-Wong, The Economics of Fog Computing: Pricing Tradeoffs for Distributed Data Analytics, *Fognet and Fogonomics, Wiley*, in print, 2019.
- T. Chang, L. Zheng, M. Gorlatova, C. Gitau, C. Huang, M. Chiang, Demo: Decomposing Data Analytics in Fog Networks, *ACM SenSys'17*, Delft, Netherlands, Nov. 2017.
- **IEEE 1934** Fog Computing Standard, 2018.

Characterizing Fog: Obtaining Quantitative Understanding

- Problem:
  - Quantitative performance characterizations of fog systems are currently lacking
- Goals:
  - Understand properties of fog execution points and options
  - Inform task placement and decomposition algorithms

- H. Inalekin, M. Gorlatova, M. Chiang, Virtualized Control over Fog: Interplay between Reliability and Latency, accepted with minor revisions to the *IEEE Internet of Things Journal*, 2018.
- M. Gorlatova, H. Inalekin, M. Chiang, under double-blind review, 2018

Characterizing Fog: Execution Points

Local execution

Server-, consumer- grade

On, off campus

Cloud execution

AWS

EC2, λ

Microsoft Azure

OR

VA

Tokyo, JP

- Public cloud:
  - Processor, network sharing
  - **Serverless execution**

Characterizing Fog: Setup and Benchmarks

Flask
web development, one drop at a time

gunicorn

- Stressing different elements
  - Compute, networking, storage
  - Multiple **complexity levels**
- Measure: all components of response times (communications, computing)

2,000+ hours of measurements

- Will make available online

CRAWDAD

Expected Tradeoffs Observed

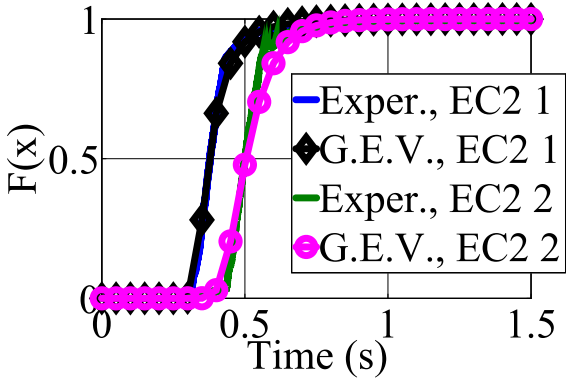
- Edge is faster than the cloud **up to a certain level of task complexity**

- Connections to the cloud are notably **faster on-campus** than in nearby residential areas

On, off campus

Suitable Execution Latency Model: Generalized Extreme Value Distribution

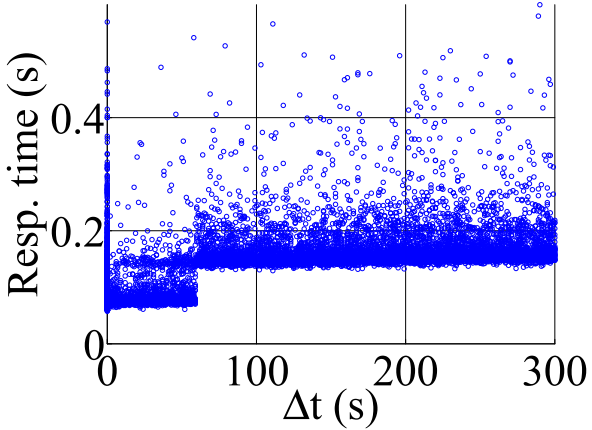
Response time CDF: AWS EC2

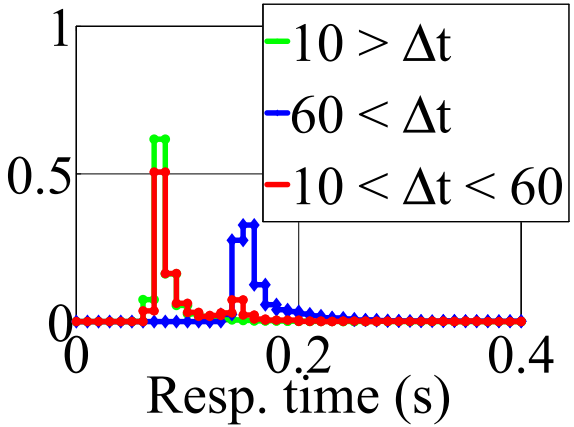


- Additional properties: CDF stability, ease of obtaining
- Limitation: more complex tasks may have execution latency distributions driven by program flow variations

Serverless Execution: Properties

- Available since 2014
- **Infinite-capacity** execution options
- Auto-scaling and spin-down
  - ➢ Response time depends on inter-invocation times



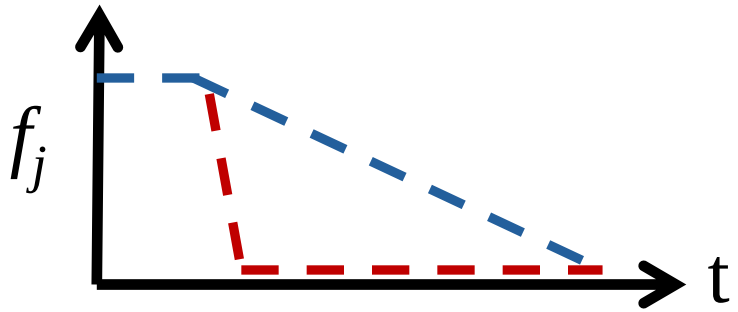


Quality-Latency Co-optimization in Execution Point Selection

- Inspired by **anytime algorithms**
  - A_{jx}: intrinsic utility of executing task j with option x
  - $f_j(t)$: utility obtained when task j completes in time t
    - Also appear in age of information approaches

$$\max U_{jx} \equiv \underset{t}{E}(A_{jx} f_j(t))$$

- Subject to capacity restrictions



Ongoing Work and Next Steps

- Design of fog task allocation and restructuring algorithms

amazon

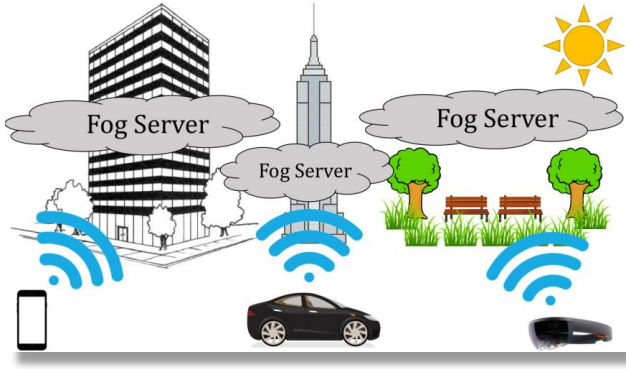
- Large-scale study of fog **latency** and **reliability**
  - ➢ In collaboration with Duke Office of Information Technology

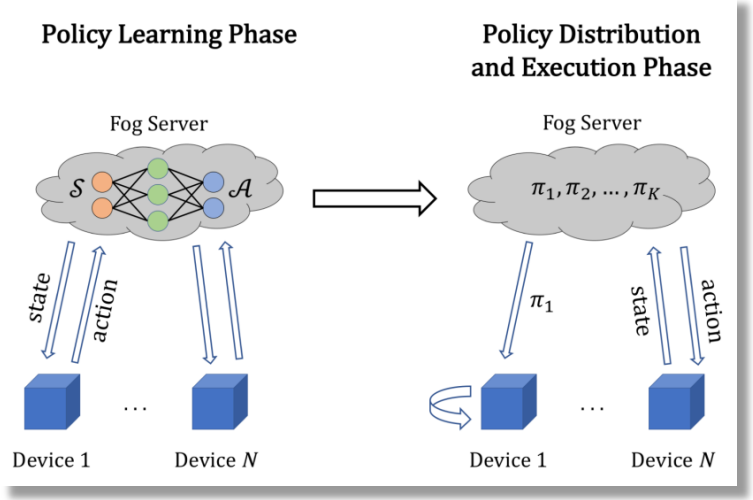
Duke OIT

- ❑ H. Inalekin, M. Gorlatova, M. Chiang, Virtualized Control over Fog: Interplay between Reliability and Latency, accepted with minor revisions to the *IEEE Internet of Things Journal*, 2018.
- ❑ M. Gorlatova, H. Inalekin, M. Chiang, under double-blind review, 2018

Towards Intelligence on the Edge

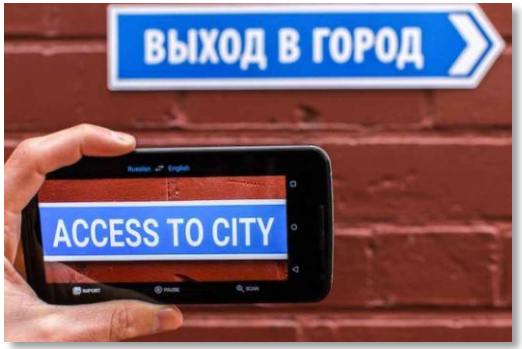
- Edge/fog computing
- Characterizing fog substrates
- New capabilities enabled by fog: intelligent augmented reality





Augmented Reality (AR): A Definition

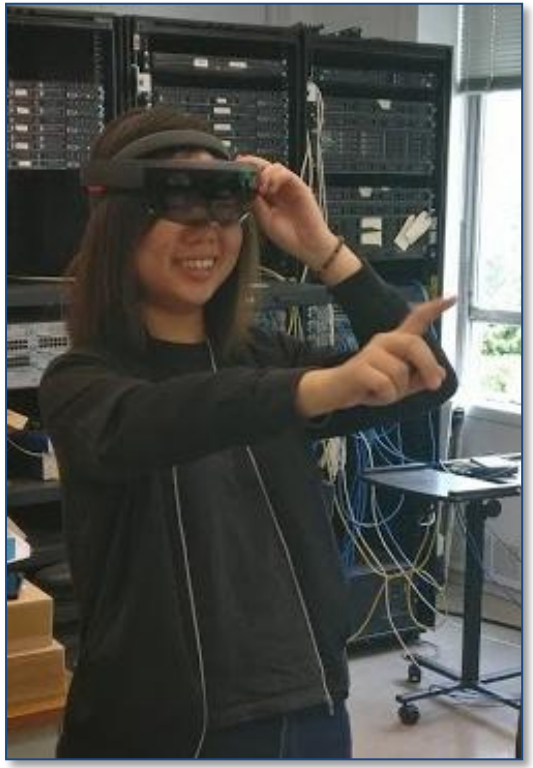
- The [virtual] content is laid out around a user **in the same spatial coordinates as the physical objects surrounding her/him***



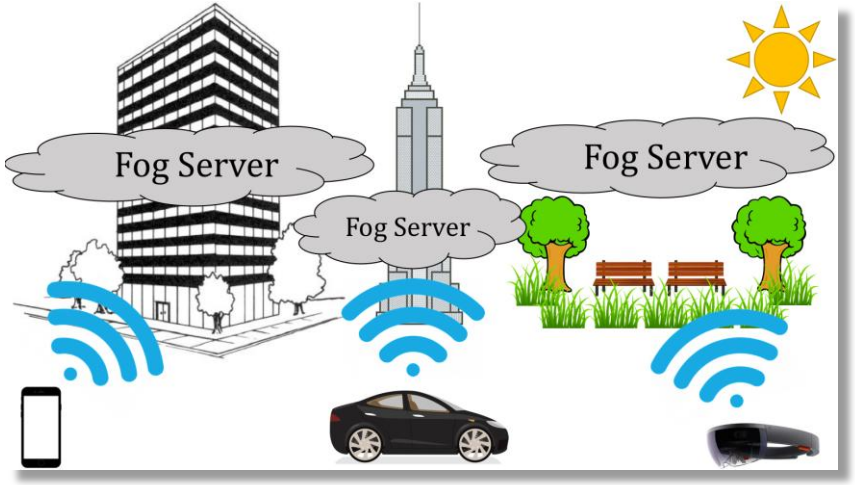
*From: Baldassi et al, Challenges and New Directions in Augmented Reality, Computer Security, and Neuroscience, June 2018.

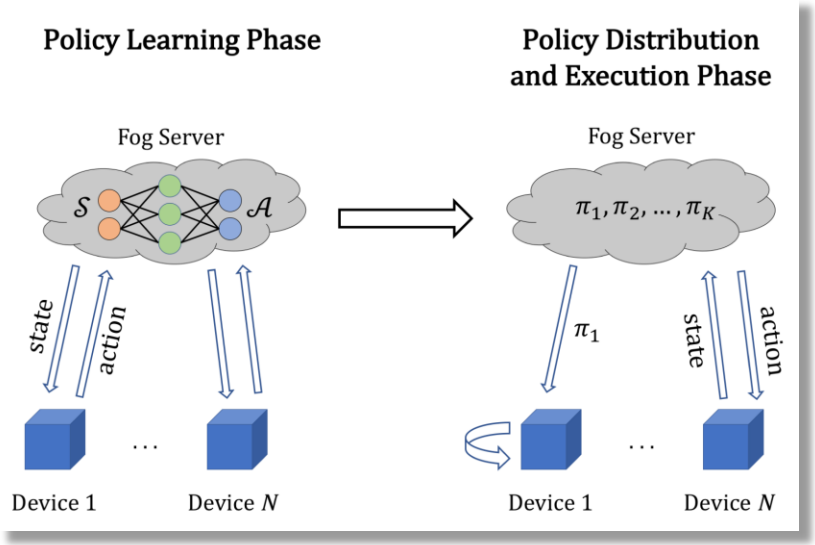
AR: Current State And Limitations

- Already: impressive immersive experiences
- Next: enabling **practical mainstream AR** with edge/fog
  - Set size reduction
  - Interactive multi-user experiences
  - **Intelligent behavior**



Fog/Edge in Support of Intelligent Augmented Reality





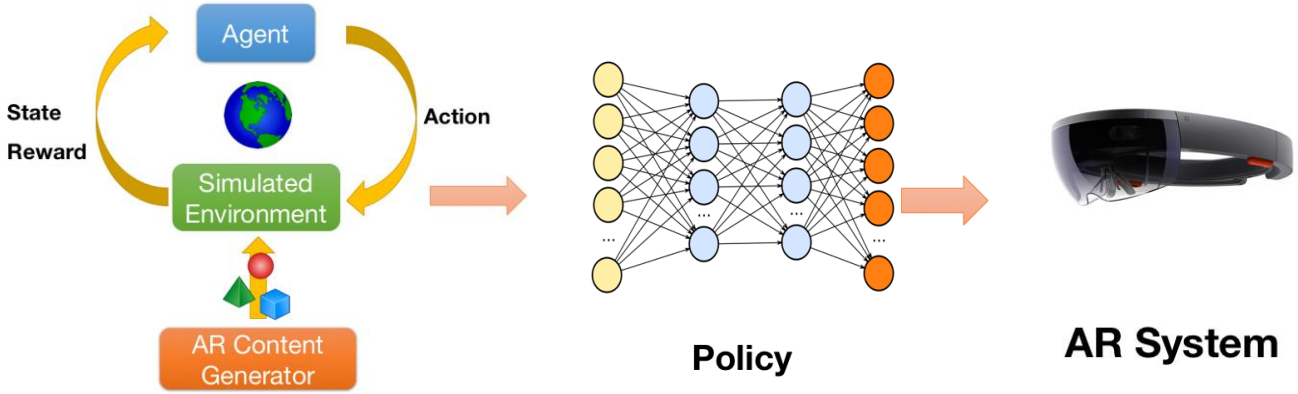
- S. Ahn, M. Gorlatova, P. Naghizadeh, M. Chiang, P. Mittal, Adaptive Fog-based Output Security for Augmented Reality, in Proc. *ACM SIGCOMM VR/AR Network Workshop*, Aug. 2018.
- S. Ahn, M. Gorlatova, P. Naghizadeh, M. Chiang, under double-blind review, 2018

Intelligent Edge for Securing AR Experiences

- Dangerous to block the view with holographic content
- State of the art: manually pre-specified fixed policies

Edge-Aided Approach: Automatic Generation of Security Policies

- With reinforcement learning



AR Security as a Reinforcement Learning Problem

1) States: Locations of real-world objects and holograms + sizes of bounding boxes

2) Actions: Change hologram locations & transparency

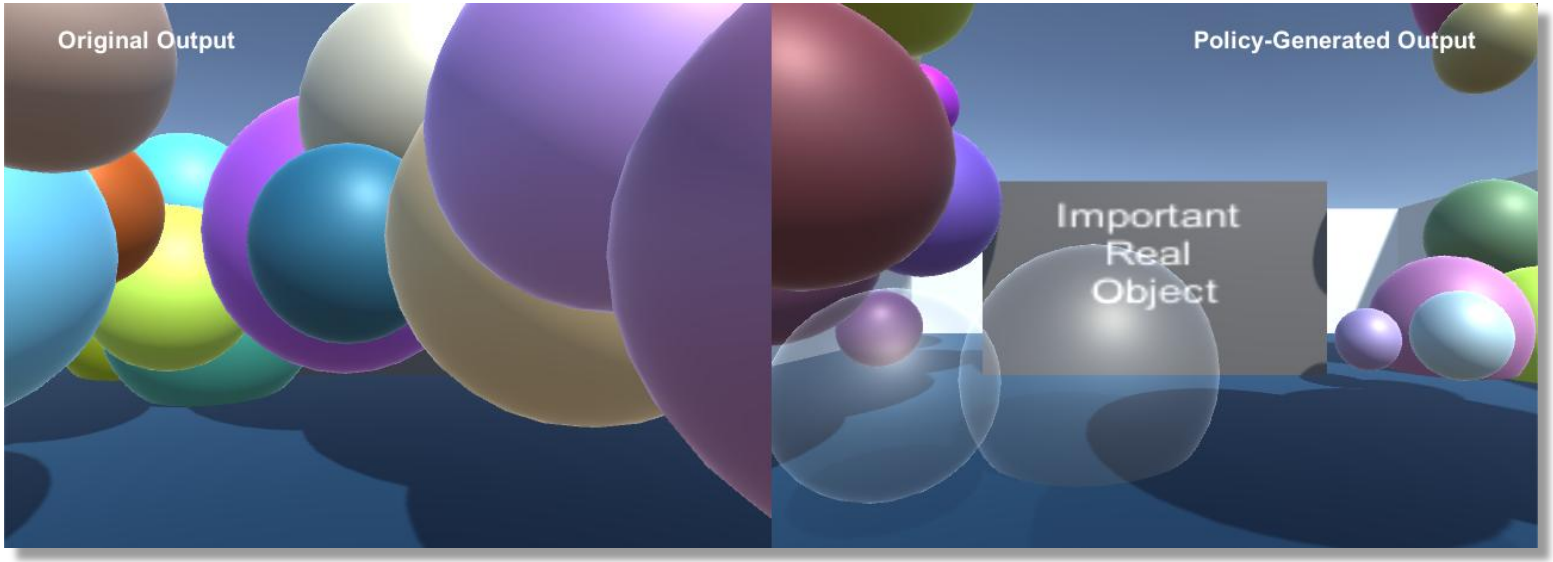
3) Reward function:

+ reward: Increasing the visibility of the real-world objects

- reward: Moving holograms far from their original position

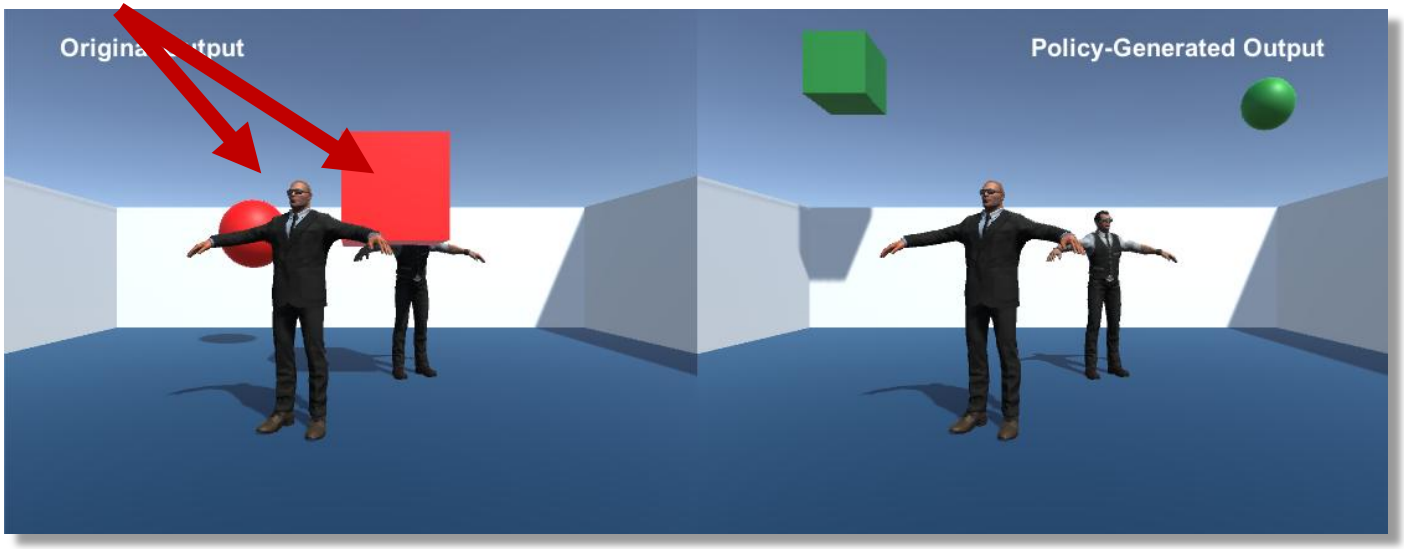
Simulation-based Training

- Unity + ML Agents SDK + TensorFlow + Proximal Policy Optimization



Edge-based Policy in Action

Moving Pedestrians



Above and beyond the state of non-edge-aided AR

Next Steps

- Intelligent AR: **behavioral cloning** for hologram positioning
- Edge-aided **multi-user AR**
- Early 2019: a pilot deployment of an intelligent edge-aided AR system on Duke University campus

- S. Ahn, M. Gorlatova, P. Naghizadeh, M. Chiang, P. Mittal, Adaptive Fog-based Output Security for Augmented Reality, in Proc*. ACM SIGCOMM VR/AR Network Workshop*, Aug. 2018.
- S. Ahn, M. Gorlatova, P. Naghizadeh, M. Chiang, under double-blind review, 2018

Towards Intelligence on the Edge: Summary

- Edge/fog computing
- Characterizing fog substrates
- New capabilities enabled by fog: intelligent augmented reality