ECE 590/COMPSI 590
Special Topics: Edge Computing

How Does Edge Help The Cloud?

Monday January 27th, 2020

Duke UNIVERSITY

1

Last Lecture: Recap

- Higher-end mobile devices
- Cloudlets
  - Current presence
  - Challenges
- Mobile offloading
- Future directions in mobile offloading

2

Duke UNIVERSITY

2

Class Outline

- Edge helping cloud
  - Why edge makes sense for the cloud
  - Background: latency and jitter
  - Challenges in supporting low-latency low-jitter solutions with modern cloud architectures
- Telecom vision for the edge
  - An infrastructure view of edge computing
  - 5G and ETSI MEC

Quiz

Research Project Timelines: A Reminder

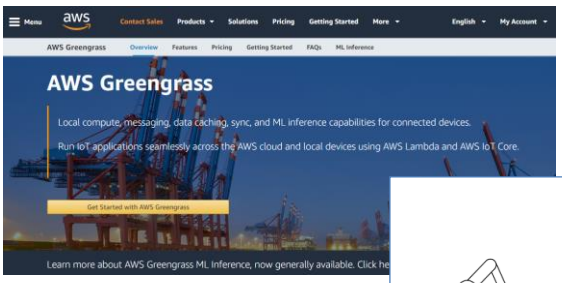
- Teams established: Friday **January 24th**
- Proposal due: Monday **February 10th**
- Progress report due: **Friday March 20th**
- Final presentations: **weeks of March 29th, April 5th, and April 12th**
- Final report due: **Friday April 17th**

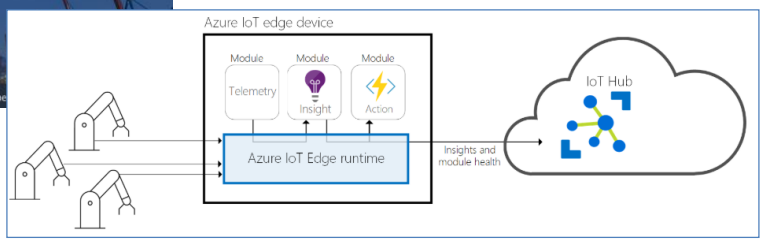
5

Duke UNIVERSITY

5

Why do Amazon and Microsoft Want to Create Edge Services?





6

Duke UNIVERSITY

6

And Why Do Telecom Giants?

7

Duke UNIVERSITY

7

Class Outline

- Edge helping cloud
  - Why edge makes sense for the cloud
  - Background: latency and jitter
  - Challenges in supporting low-latency low-jitter solutions with modern cloud architectures
- Telecom vision for the edge
  - An infrastructure view of edge computing
  - 5G and ETSI MEC

8

Duke UNIVERSITY

8

Why do Amazon and Microsoft Want to Create Edge Services?

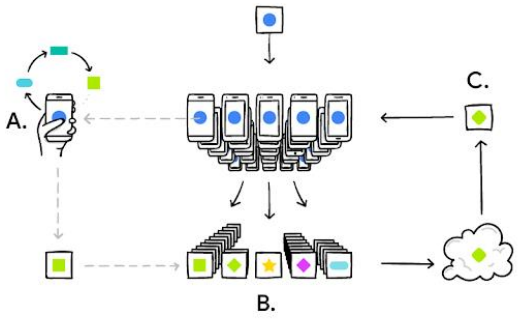
- First of all, not to be left out of the game
  - Most likely, you **will** have an IoT gateway, and you will run **something** on it

Challenges in Cloud Interacting with IoT Nodes

- Some similar to serverless computing
  - Short requests from billions of devices
  - Difficult to right-size resources

Special Case: Cloud Company Owning the Datacenter and the Workloads

- Federated learning example
- Letting local devices do the work
  - Datacenter does not have to
- Most likely not the primary motivation



11

Duke UNIVERSITY

11

Fundamental Technical Reason: Challenges in Supporting Low-Latency Services

- Come up in context of existing latency-sensitive services
  - Responsive applications
  - Distributed data analytics

12

Duke UNIVERSITY

12

Class Outline

- Edge helping cloud
  - Why edge makes sense for the cloud
  - Background: latency and jitter
  - Challenges in supporting low-latency low-jitter solutions with modern cloud architectures
- Telecom vision for the edge
  - An infrastructure view of edge computing
  - 5G and ETSI MEC

13

Latency Components

- Latency, in a distributed system:
  - Getting data to and from the execution point
  - + service invocation time
  - + service execution time

14

Latency with Edge and Cloud

- Cloud:
  - ➢ Globally **pooled** users → central server farm
- Edge:
  - ➢ Local users → local gateway/cloudlet

Latency with Edge and Cloud: Comparison (1/2)

- Cloud communication latency strictly greater than edge latency
  - ➢ Speed of light

From:http://ipnetwork.bgtmo.ip.att.net/pws/network_delay.html

U.S. Network Latency

Figures are in **ms**. Thresholds are distance sensitive.

Current Overall Average: 31 ms

CITY PAIRS	Atl	Aus	Cam	Chi	Cle	Dal	Den	Det	Hou	Ind	Kan	LA	Mad	Nas	NO	NY	Orl	Pa	Phx	SA	SD	SF	StL	Sea	Was
Austin	27																								
Cambridge	28	52																							
Chicago	24	30	24																						
Cleveland	18	35	19	7																					
Dallas	17	6	50	26	27																				
Denver	37	25	42	19	26	19																			
Detroit	21	39	21	7	4	34	26																		
Houston	18	7	44	30	34	5	24	39																	
Indianapolis				5	8																				
Kansas City	22	15	31	12	18	10	15	18	15																
Los Angeles	46		67	44	50	28	25	51	33		39														
Madison				5							18														
Nashville	7	22	28	17	11	19	33	14	25		17	44													
New Orleans	12	13	37	33	31	12	32	29	7		21	40		20											
New York	24	49	6	18	13	45	37	17	40		26	61		23	34										
Orlando	11	27	37	33	28	24	43	30	19		33	53		15	14	32									
Philadelphia	22	43	9	17	10	40	36	15	39		25	59		21	32	3	32								
Phoenix	40	20	64	42	45	20	34	49	23		29	10		35	29	59	43	56							
San Antonio	26	3	51	32	33	8	25	40	5		17	28		26	11	48	25	43	18						
San Diego	43	27	68		54	26	28	53	29		35	4		47	35		49	63	7	25					
San Francisco	51	41	67	45	51	37	28	52	40		43	8		60	48	62	60	61	19	37	13				
St. Louis	18	22	26	7	13	17	21	15	23	13	6	45	11	11	29	21	27	19	37	24	44	50			
Seattle	64	54	69	41	50	48	30	48	53		44			57	60	65		62	37	55	31	16	53		
Washington	18	38	10	20	13	33	38	17	37		22	61		23	29	5	27	4	53	40	59	63	16	60	

Latency with Edge and Cloud: Comparison (2/2)

- Cloud **communication** latency:
  - Affected by complex underlying global networking infrastructures
    - Multiple hops, multiple switches in the way
- Cloud **execution** latency:
  - Can be smaller than edge latency
  - Affected by complex datacenter sharing mechanisms

Providing latency **guarantees** is a challenge for the cloud

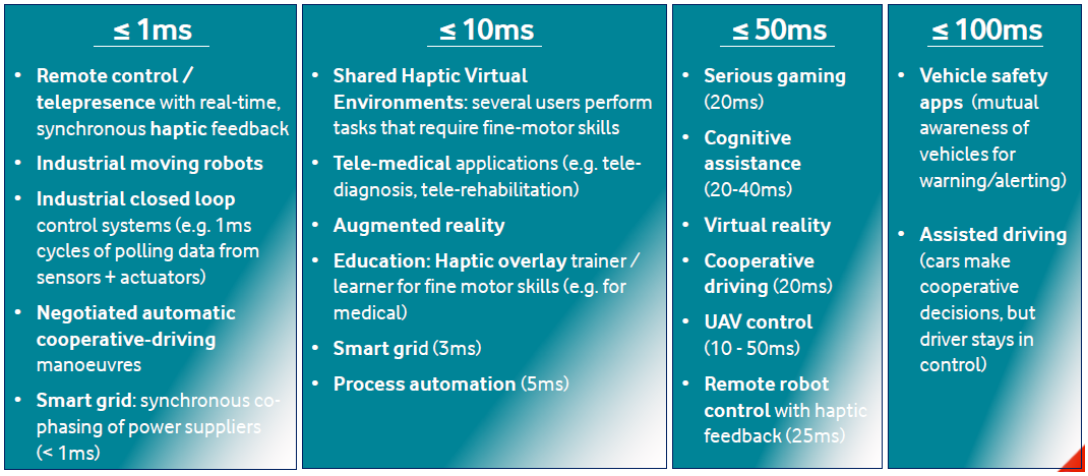
17

Latency Requirements (1/2)

- Web world's take on latency:
  - Goes back to late 1960s work by Miller et al, on *response time in man-computer conversational transactions*
  - **100 ms** for a fluid computer response feeling
  - Loss of user attention after **5-10 s**
- Web queries are optimized for 100 ms latency

18

Latency Requirements (2/2)



From: Simone Mangiante, Through the Fog Workshop, Feb. 2017

19

Duke UNIVERSITY

19

Mean Latency and Jitter Both Matter

- Jitter: deviations from the mean
- Jitter is problematic for voice, gaming, video conferencing, control, augmented reality, …

20

Duke UNIVERSITY

20

Class Outline

- Edge helping cloud
  - Why edge makes sense for the cloud
  - Background: latency and jitter
  - Challenges in supporting low-latency low-jitter solutions with modern cloud architectures
- Telecom vision for the edge
  - An infrastructure view of edge computing
  - 5G and ETSI MEC

21

Cloud Latency: Background

- Recognize latency magnitude as an issue
  - E.g., Content Delivery Networks as one solution
- Recognize jitter as an issue
  - E.g., for multi-player games, VoIP
    - Edge should be able to support applications with tighter latency requirements

22

Cloud Providers Viewpoint

- Client-specific latency performance requirements are difficult to satisfy
- Hide the details of the underlying infrastructure
  - Can evolve it without getting locked into outdated design decisions
  - Avoid revealing trade secrets

From: Inferring the Network Latency Requirements of Cloud Tenants, Mogul et al, USENIX HotOS'15

23

Latency Variability Sources (1/3)

- Shared Resources
  - CPU cores
  - Processor caches
  - Memory bandwidth
  - Network bandwidth
- In our measurements with AWS t2.micro, we have seen up to 11x increase in latency

From: The Tail at Scale, J. Dean et al, Communications of the ACM, 2013

24

Latency Variability Sources (2/3)

- Daemons
- Global resource sharing, across multiple machines
  - Network switches, shared file systems
- Maintenance activities
  - E.g., log compaction

From: The Tail at Scale, J. Dean et al, Communications of the ACM, 2013

25

Latency Variability Sources (3/3)

- Queuing
  - Intermediate servers, network switches
- Power limits
  - Throttling if power envelope is exceeded for a long time
- Energy management
  - Latency when moving from inactive to active states

From: The Tail at Scale, J. Dean et al, Communications of the ACM, 2013

26

There are Ways of Improving Cloud Latency Support

- E.g.,
  - For latency caused by shared network or CPU: isolated resources
- But:
  - All require additional resources
  - New applications need even tighter latencies

27

Possible Future Combined Edge-Cloud Architecture

- Latency-oriented reservation-based solutions on the edge
- Traditional sharing-oriented solutions on the cloud

28

Summary: Why Edge Makes Sense for the Cloud

- Capturing new business opportunities
- Overcoming IoT node management complexity
- Solving latency challenges

Class Outline

- Edge helping cloud
  - Why edge makes sense for the cloud
  - Background: latency and jitter
  - Challenges in supporting low-latency low-jitter solutions with modern cloud architectures
- Telecom vision for the edge
  - An infrastructure view of edge computing
  - 5G and ETSI MEC

Telecom Providers (1/2)

- Phone, internet, TV
- Mobile wireless service

31

Telecom Providers (2/2)

Company	Country	Market value ($ Bn)	Revenue	Profit
China Mobile	China	213.8	88.8	20.5
AT&T	USA	200.1	127.3	7.3
Verizon Communications	USA	137.3	115.7	0.9
Vodafone	UK	135.7	74.4	11.1
América Móvil	Mexico	70.7	60.2	7.1
Telefónica	Spain	67.1	82.3	5.2
Telstra	Australia	58.4	25.8	3.5
Nippon Telegraph & Tel	Japan	58.2	127	5.6
Deutsche Telekom	Germany	48.8	76.7	-7
Softbank	Japan	47.2	38.78	3.8

From: https://en.wikipedia.org/wiki/Telecommunications_industry

32

Interest in Edge

Wireless

AT&T: 2020 will be year of the edge

by Monica Alleven | Jan 8, 2020 5:02am

Telefónica makes TV companies' dreams come true with edge computing

05 DECEMBER 2019

Verizon and AWS announce 5G Edge computing partnership

Brian Heater @bheater / 2:53 pm EST • December 3, 2019

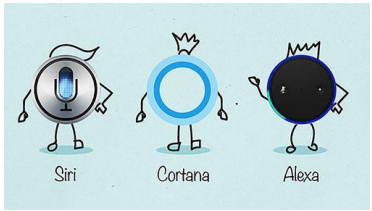
33

Class Reading Materials Question

- Did you find the perspective on edge computing, as outlined in the Vodafone presentation, different from what we have been discussing?

34

Mobile Offloading: Application View



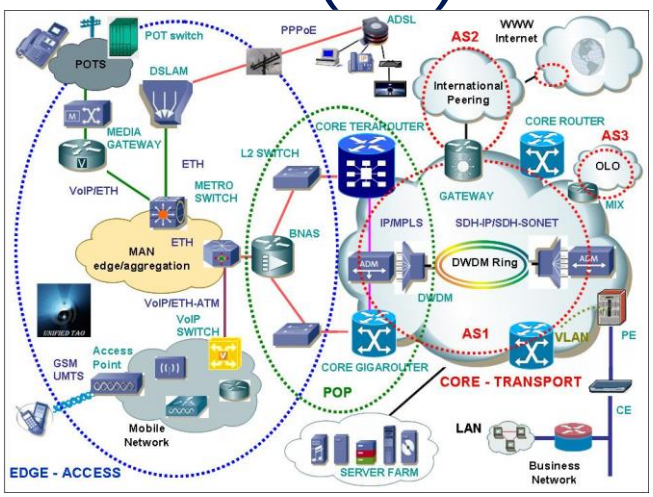
- The view we have seen so far
- But, there is telecom piping underneath **all** of it

35

Duke UNIVERSITY

Mobile Offloading: Infrastructure View (1/2)

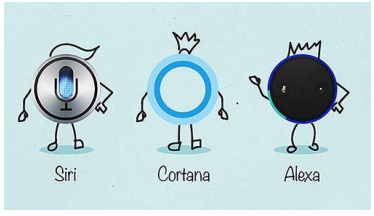
Network diagram from :
http://unifiedtao-en.blogspot.com/2012/08/from-complicated-to-complex-to.html

- 

36

Duke UNIVERSITY

Mobile Offloading: Infrastructure View (2/2)



- Infrastructure:
  - Pervasive
  - Expensive
    - Including real estate, laying and maintaining wires, …
  - Mission-critical
- Over-the-top content (**OTC**) providers have it easy, in comparison

37

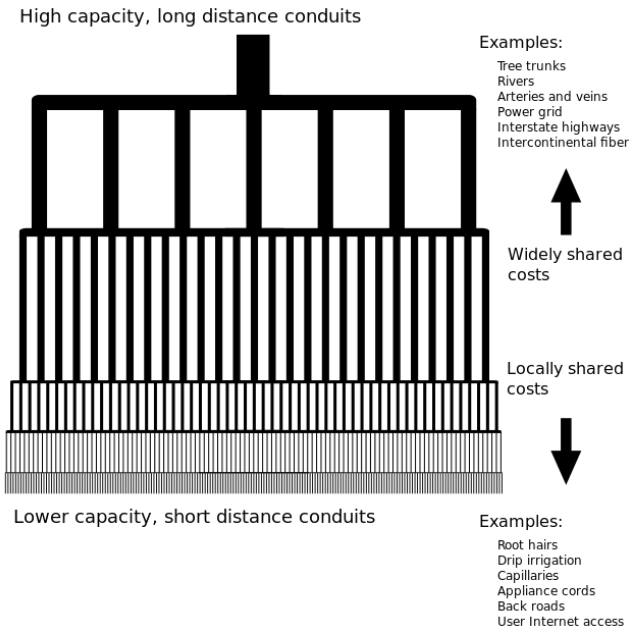
Telecom as an Infrastructure Layer

- Telecom as a utility
  - Commoditization of telecommunication services
  - "Metered data" services, minutes of voice, number of texts
  - Hard to differentiate offerings from different companies
- Connectivity services → connected experiences
  - Not exclusive to edge services
  - … but very important in edge context

38

Last-Mile Delivery is Expensive

- Edge-based data processing can help



39

Duke UNIVERSITY

39

Class Outline

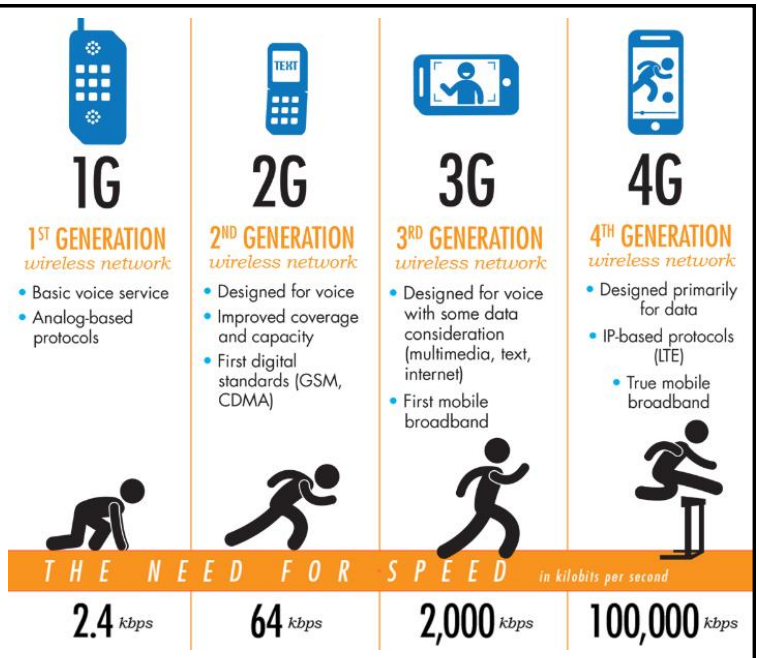
- Edge helping cloud
  - Why edge makes sense for the cloud
  - Background: latency and jitter
  - Challenges in supporting low-latency low-jitter solutions with modern cloud architectures
- Telecom vision for the edge
  - An infrastructure view of edge computing
  - 5G and ETSI MEC

40

Duke UNIVERSITY

40

Cellular Technologies Have Been Evolving

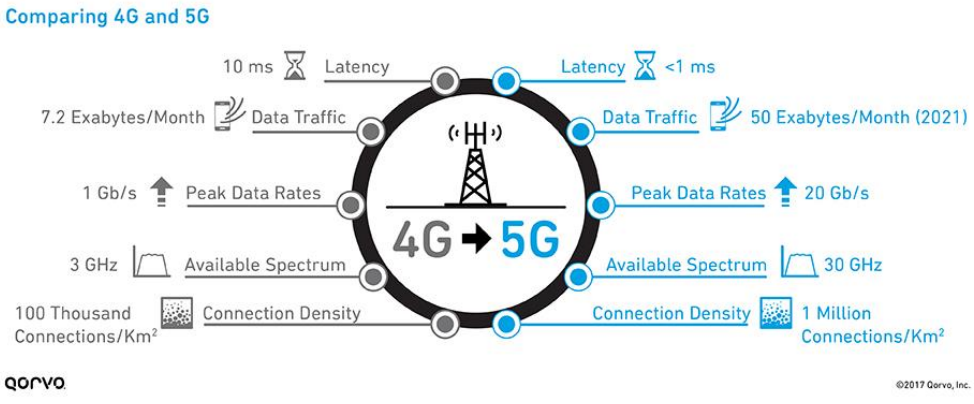


Duke UNIVERSITY

41

5G is Coming

- Pilot deployments, trials ongoing



42

Duke UNIVERSITY

42

Edge Computing is a Part of 5G

- One of the building blocks
- Offers:
  - Lower latency
  - Reduced load on core network
- Idea: co-locate edge computing servers with cellular base stations

43

ETSI MEC (1/2)

- Standardization effort:
  - European Telecommunications Standards (ETSI) **Multi-access Edge Computing** (MEC)
- Since 2014

44

ETSI MEC (2/2)

- Many participating companies

45

ETSI MEC: Example Standards

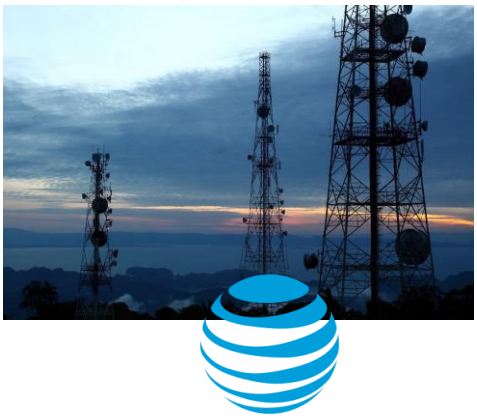
- Study on MEC support for V2x use cases
- UE identity API
- System, host, and platform management
- Bandwidth management API
- UE application interface
- Application lifecycle, rules and requirements management
- Radio Network information API
- Location API
- …

46

Can Better Take Advantage of Existing Infrastructure

- Present in your zip code
- Present in your house
  - In contrast to cloudlets

COMCAST



47

Recall: Cloudlet Challenges

- **Mobile** devices → supporting mobility
  - No related concepts in cloud computing

48

Telecom Edge vs. Cloudlet Edge (1/2)

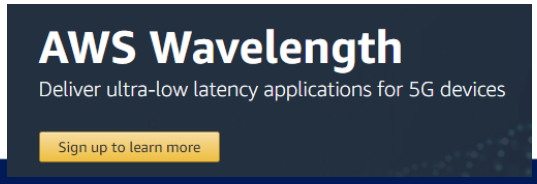
- Existing pervasive infrastructure
- Minimal possible latency for cellular devices
- Know **all** about mobility
  - ➢ Have a concept of location – can geo-locate without a GPS
  - ➢ Know how to handle handoff
    - However, computing handoff ≠ wireless hand-off

Telecom Edge vs. Cloudlet Edge (2/2)

- Different mentality than Amazon, Microsoft, Google
  - ➢ Reliability-oriented
  - ➢ Slow to change
    - Standards rather than iterative deployments
  - ➢ Far less experience in creating developer ecosystems
- For academic work: difficult to test your ideas

Recent Development: Partnerships with Amazon

- Verizon, Vodafone partnering with AWS on an AWS Wavelength execution
  - Write AWS applications, execute in local "Wavelength Zones"
  - Currently in private beta in limited markets



51

Duke UNIVERSITY

51

Recap

- Edge helping cloud
  - Why edge makes sense for the cloud
  - Background: latency and jitter
  - Challenges in supporting low-latency low-jitter solutions with modern cloud architectures
- Telecom vision for the edge
  - An infrastructure view of edge computing
  - 5G and ETSI MEC

52

Duke UNIVERSITY

52

Next Class

- Augmented and Virtual Reality on the Edge
- ML on the Edge

Next Class: Homework

- Reading for the class
  - Google AI Blog: Federated Learning
- Work on your research project