ECE 590/COMPSI 590
Special Topics: Edge Computing

Edge Helping Higher-end Mobile Devices: Mobile Offloading

Wednesday September 5th, 2018

Duke UNIVERSITY

Last Class Recap

- Edge and IoT devices
  - Common IoT architectures
  - Role of the gateway
- Opportunities: edge for responsive IoT applications
  - Hardware
  - Algorithms
  - Edge for system decisions

2

Duke UNIVERSITY

Quiz

Lecture Outline

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

Edge for IoT Nodes vs. Edge For High-End Mobile Nodes

Low-end IoT nodes

IoT nodes → Gateway → Cloud

Last lecture

Higher-end mobile devices

Mobile device→ Cloudlet → Cloud

This lecture

5

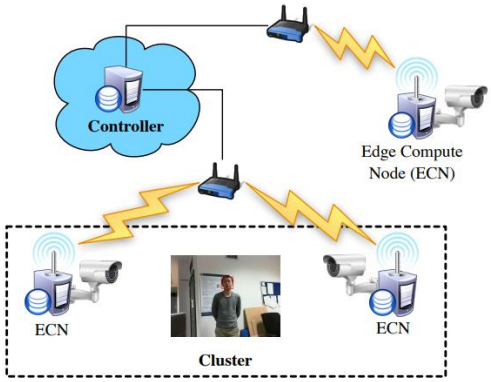
Higher-End Mobile Devices

- Mobile phones: prevalent use case
- AR/VR, drones, smart cars – emerging use cases

6

Special Case: Camera Installations

- E.g., city, campus security cameras
  - Very common
  - Of major practical importance
  - Often not **mobile** devices
  - Many video-specific mechanisms
- Time-permitting, will cover later in the course



From: The Design and Implementation of a Wireless Video Surveillance System, Zhang et al, ACM MobiCom'15

7

Unlike IoT devices…

- Not as resource-constrained
  - Fewer per-device customizations
  - Usually standard protocols



8

Unlike IoT devices…

- Complex, often high-volume, data
  - Variety of sensors – accelerometers, video, audio, …
- More complex operations
  - Thinking in full application pipelines, rather than individual tasks

9

Duke UNIVERSITY

Like IoT devices…

- Battery-limited
  - How long they last
  - How much heat they produce
- Usability limited by the batteries

Duke UNIVERSITY

Mobile Device vs. a Server

- Isn't a mobile device a desktop in your pocket?
- Server > mobile device
  - Power constraints → 500 W of power on a high-end GPU, 10 W on a mobile SoC GPU
  - Space constraints

Cloudlets

- Local mini-clouds
- Envisioned properties:
  - Powerful, well-connected, and safe
  - Close at hand
  - Build on standard cloud technology

Existing and Possible "Cloudlets"

- On-site computing
- Targeted edge installations
- Resource scavenging

On-Site Computing

- At universities
- … and other medium and large organizations
  - Shrinking but not disappearing
  - Usually have low utilization

Targeted Installation: Chick-fil-A (1/2)

July 2018

15

Duke UNIVERSITY

Targeted Installation: Chick-fil-A (2/2)

Azure

AWS

Google Cloud

Cloud-fil-A

16

Duke UNIVERSITY

Can Imagine Deploying More of These

- Especially for Augmented and Virtual Reality
- … and for smart cities in general

17

Duke UNIVERSITY

Resource Scavenging

- Finding unused devices around you
  - "Cyber foraging"
  - "*Uberization*" of computing and storage

18

Duke UNIVERSITY

Resource Scavenging: Open Questions

- Open questions
  - Discovery, connectivity
  - Security
  - Incentives, pricing

- Smart city computing infrastructures, e.g. Barcelona deployments, try to address these

19

Cloudlet Challenges

- **Mobile** devices → supporting mobility
- Cloud**let** →does not have the scale of the cloud

20

Cloudlets Helping Mobile Devices: Challenges: Rapid Service Provisioning

- A scenario: a student comes to Hudson Hall and needs to use our cloudlet
  - Service discovery
  - Provisioning delay
  - Do not have the scale of the cloud: do we prioritize this user over others? Shift workloads with every user?

21

Cloudlets Helping Mobile Devices: Challenges: Service Handoff

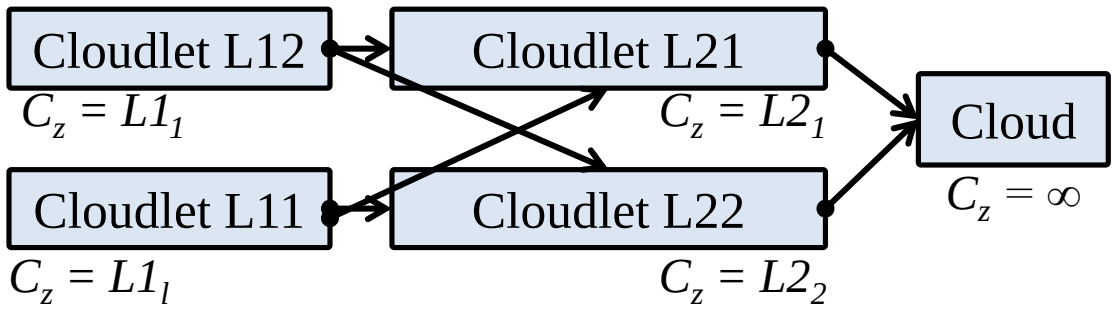
- A scenario: the student moves from Hudson Hall to CIEMAS
  - Do we transfer their workload state?
  - Do we de-provision their Hudson Hall services?

22

Cloudlets Helping Mobile Devices: Challenges

- Platform challenges
  - Challenges similar to wireless hand-off
- Workload allocation and scheduling challenges

Workload Allocation and Scheduling Challenges



- Settings become complicated very quickly
  - Cloudlet capacities
  - Network restrictions

Our Recent Work: Picking Workloads for Local Execution (1/2)

- Work joint with  
- Workloads achieve different utilities when executed in different locations
  - Quality
  - Latency
- Want to maximize the sum of utilities for all workloads

25

Our Recent Work: Picking Workloads for Local Execution (2/2)

- Easy to solve for one capacitated cloudlet
  - Pick the workloads with the highest utility gains over the cloud
- Becomes complex with as few as two capacitated cloudlets

26

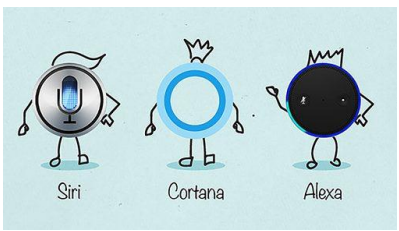
Lecture Outline

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

27

How Edge Helps: Mobile Offloading

- Executing code **not** on the mobile device
- E.g., image, video, audio, other sensor data processing
  - Face detection, person identification
  - Language translation, speaker identification
  - Activity tracking, gesture recognition



- All offload processing to the cloud

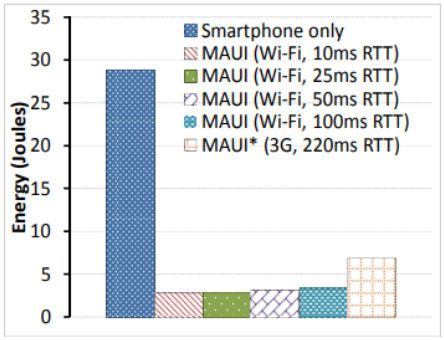
28

Goals: Reducing Mobile Device Energy Consumption (1/3)

- Need to have:
  - Energy to {transmit data + receive results} < energy to {execute the operation on the mobile device}
- Design principles:
  - Pick the most compute-intensive parts of the operation
  - Reduce the size of what is transmitted: data and results
- Order-of-magnitude mobile energy savings possible

29

Example: Face Recognition with MAUI



ONE RUN FACE RECOGNITION

From: MAUI: Making Smartphones Last Longer with Code Offload, Cuervo et al., ACM MobiSys'10.

30

Goals: Reducing Mobile Device Energy Consumption (2/3)

- Not minimizing **total** energy:
  - Combined server + mobile energy spending can be higher than mobile-only energy spending
- System heterogeneity principle:
  - Server energy spending is not as important as mobile device energy spending
  - Server grade does not factor into energy minimization objective

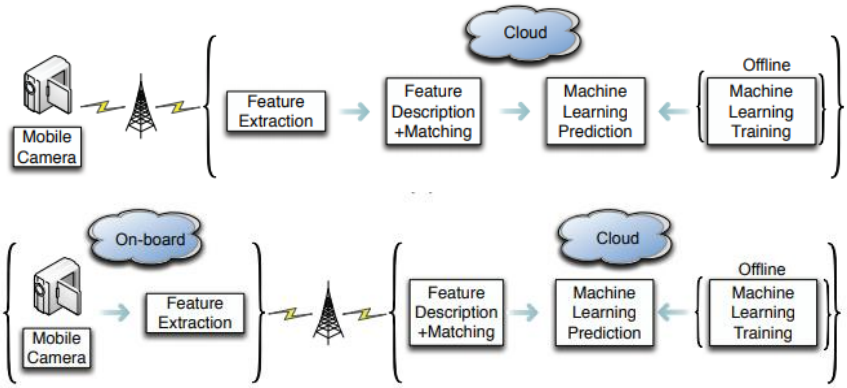
31

Goals: Reducing Mobile Device Energy Consumption (3/3)

- Often: transmit partially processed, rather than raw, data
  - Energy to {extract features + transmit extracted features + receive results} < energy to {transmit data + receive results}
  - Energy to {extract features + transmit extracted features + receive results} < energy to {execute the operation on the mobile device}

32

Local Feature Extraction Can Reduce the Amount of Data Transmitted



From: A Hybrid Approach To Offloading Mobile Image Classification, Hauswald et al, IEEE ICASSP'14.

33

Goals: Minimizing Task Completion Time

- Need to have:
  - Time to {transmit data + execute operation on the server + receive the results} < time to {execute the operation locally}
- More demands on the server:
  - Need to offload to a much more capable device

34

Mobile Offloading: Need for Scheduling Mechanisms

- Time, energy vary with network connectivity
- Need to make decisions for different conditions
  - Different ways of placing different parts of operations
  - Offline versus online
  - Joint scheduling of different operations
  - Scheduling that takes into account different local processors and the cloud

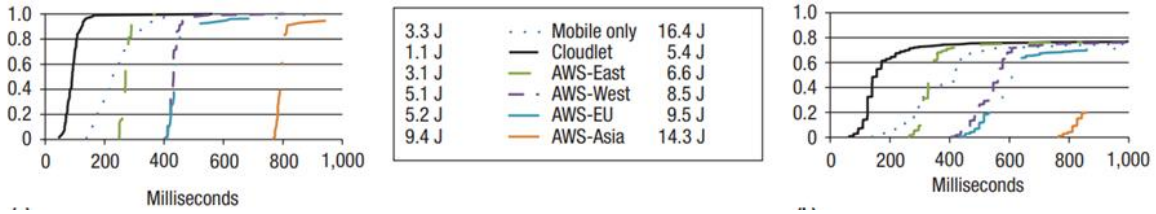
35

Role of the Edge (1/2)

- Short transmission distance helps both transmission energy and latency
  - Better performance of existing offloading scenarios
  - Offloading equations “work out” in more cases
- Potentially, additional privacy

36

Role of the Edge (2/2)



Augmented reality

Face recognition

From: The Emergence of Edge Computing, by M. Satyanarayanan, IEEE Computer, 2017. Adapted from The Impact of Mobile Multimedia Applications on Data Center Consolidation, by Ha et al, 2013.

Lecture Outline

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

Future Directions: "Offload Shaping"

- Adapting operations for offloading
- A form of creative pre-processing
  - Changing application pipelines **specifically** for offloading
- Some examples from: The Case for Offload Shaping, by Hu et al, ACM HotMobile'15

39

Offload Shaping: Object Recognition in Video Captures (1/2)

- Object recognition works poorly on blurry frames
  - Can drop blurry frames before transmitting them to the cloud/cloudlet for processing

	Send all	Drop blurry
Bytes transferred	0.51M	0.34M
Glass energy (J)	429(2)	292(3)
Server CPU usage (normalized)	1.00(0.01)	0.81(0.01)

40

Offload Shaping: Object Recognition in Video Captures (2/2)

- Results from similar frames are likely to be the same
  - Discard frames that are sufficiently similar

	No shaping	Drop similar	Improve-ment
Bytes transferred	0.51M	0.23M	55%
Frames recognized	171(2)	189(1)	11%
Glass power (W)	1.82(0.01)	1.83(0.01)	-1%
E2E latency (ms)	920(8)	393(2)	57%
Glass energy (J/frame)	1.66(0.01)	0.72(0.01)	57%
Server CPU usage (normalized)	1.00(0.01)	0.27(0.01)	73%

41

Offload Shaping

- (+) Holistic view of the entire system
  - Fixing inefficiencies that become obvious when we think about the system beginning-to-end
- (-) Solutions likely to be application-specific
  - E.g., blur detection in one of the previous examples

42

Opportunities: Providing Local Context

- Information about local conditions
  - Pre-programmed
  - … or learned
- Historic data, predictions

Opportunities: Providing Local Context

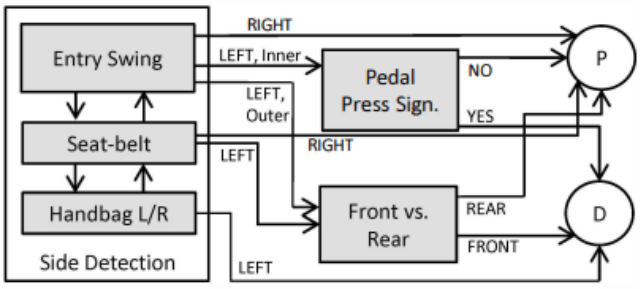
- Especially when context is large
- Opportunities for behavior **specialization**

Side Note: Context Awareness in Applications is Not New

- Traces back to early 1990s
- E.g.:
  - Active badge location system
  - OS updates only when a phone is plugged in and is on WiFi

Context Awareness Example: Is a Cell Phone User a Driver or a Passenger?

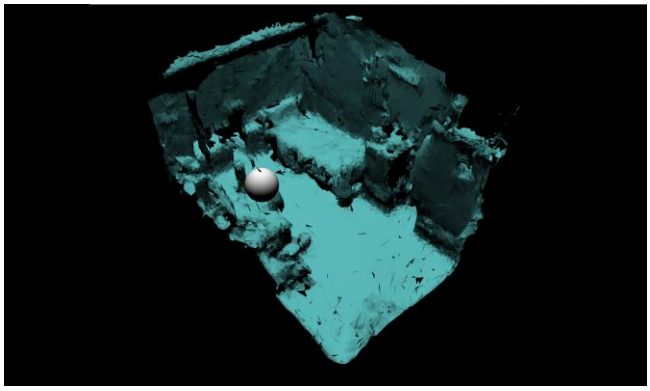
- Using cell phone's gyroscope, accelerometer data to detect and differentiate specific motions



From: I Am a Smart Phone and I Know My User is Driving, Chu et al, IEEE COMSNETS'14.

Large Local Context: 3D Maps of the Environment for AR/VR (1/2)

- Massive amounts of information and processing
  - Useful to not regenerate for all users
  - Useful to not fetch from the cloud

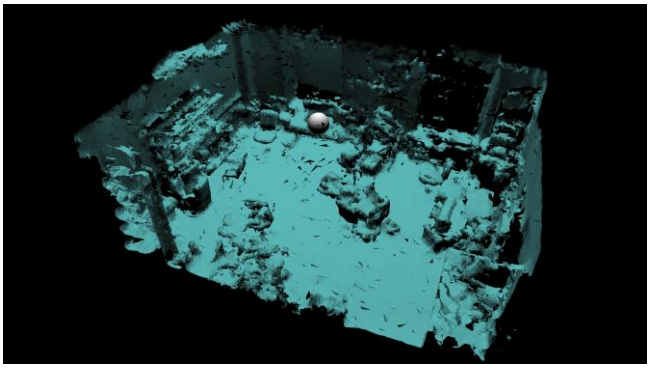


Mesh representing a student dorm room

47

Duke UNIVERSITY

Large Local Context: 3D Maps of the Environment for AR/VR (2/2)



Mesh representing a lab

48

Duke UNIVERSITY

What Could Hudson Hall and CIEMAS Cloudlets Tell Us?

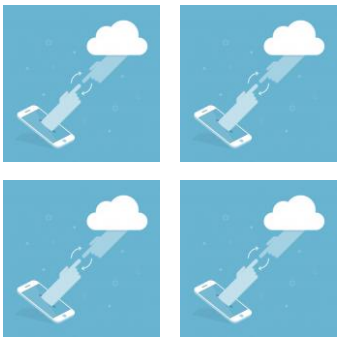
- Opportunities for behavior specialization

49

Duke UNIVERSITY

Opportunities: Thinking Across Multiple Devices and Multiple Applications

- New paradigms



- Without the cloudlets, nearby devices have no exposure to each other's actions
  - ➢ No single "choke point"

50

Duke UNIVERSITY

Opportunities: Thinking Across Multiple Devices and Multiple Applications

- Same application likely to be invoked on different devices served by one cloudlet
- Invited speaker in October

Class Recap

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

Next Class: Edge Helping the Cloud

- Why do **cloud computing companies** want to create edge services?
- Why do **telecom companies** want to create edge services?

Reading Material for the Next Class

- Technical:
  - *Commoditization of the wireless industry*
  - *Vodafone perspective on edge computing*
- Approaching research (and technology):
  - *Technology and Courage*

Homework

- Work on your research project
  - And on your proposal specifically