Edge Computing
A New Disruptive Force

Mahadev Satyanarayanan
School of Computer Science
Carnegie Mellon University
GE has more than 4000 employees focused on Edge for a broad range of applications including power, aerospace, transportation, healthcare, oil & gas, and renewables. Our objective this year is to bring together industry experts, and business thought leaders to share insights on GE products and innovations. We have 200+ GE employees attending the symposium who work on the Edge services, including energy, digital, aviation, healthcare, transportation, and oil & gas. The theme for
AWS Greengrass

Local compute, messaging, data caching, and sync capabilities for connected devices.
Run IoT applications seamlessly across the AWS cloud and local devices using AWS Lambda and AWS IoT.

Introducing IoT Edge

PREVIEW

- Run artificial intelligence at the edge
- Perform edge analytics
- Deploy IoT solutions from cloud to edge
- Manage devices centrally from the cloud
- Operate with offline and intermittent connectivity
- Enable real-time decisions
- Connect new and legacy devices
- Reduce bandwidth costs
Maverick* Research: The Edge Will Eat the Cloud

Published: 22 September 2017    ID: G00338633
How Far We Have Come!

2009 NSF Panel Summary for my Expeditions Proposal

“Many panelists do not agree with the premise of the proposal in which distant cloud computing incurs too high latency to be acceptable by mobile applications. They question the validity of such assumption as the proposal provides no real data to justify it.”

Needless to say, the proposal was rejected 😞

Time has proven the premise to be correct!

(NSF hosted workshop on “Research Challenges in Edge Computing” in 2016)
Why Is Edge Computing So Valuable?

1. Highly responsive cloud services
   “New applications and microservices”

2. Edge analytics in IoT
   “Scalable live video analytics”

3. Exposure firewall in the IoT
   “Crossing the IoT Chasm”

4. Mask disruption of cloud services
   “Disconnected operation for cloud services”

Latency (mean and tail)
Bandwidth (peak and average)
Privacy
Availability

“The Emergence of Edge Computing”
Satyanarayanan, M.
IEEE Computer, Vol. 50, No. 1, January 2017
What is a Cloudlet?

aka “micro data center”, “mobile edge cloud”, “fog node”

**Small data center at the edge of the Internet (many sizes & forms)**

- one wireless hop (+fiber or LAN) to mobile devices (Wi-Fi or 4G LTE or 5G)
- multi-tenant, as in cloud
- good isolation and safety (VM-based guests)
- lighter-weight containers (e.g. Docker within VMs) also possible

**Non-constraints** (relative to mobile devices)

- energy
- weight/size/heat

*Catalyst for new mobile applications*
Internet

Cloudlet-1
Cloudlet Services & Application Back-ends

Cloudlet-2
Cloudlet Services & Application Back-ends

Cloudlet-N
Cloudlet Services & Application Back-ends

Like a CDN for Computation
LIVING EDGE LAB
An Open and Flexible Resource for Hands-on Experience with Edge Computing

Mission Statement
“We are building a real-world testbed for Edge Computing with leading edge applications and user acceptance testing.”

Key Elements
- **Partnership:** developers for apps, services and devices join forces with telco, infrastructure and research
- **Test Diversity:** various testbeds and latest technology available for a variety of use-case scenarios
- **Open Platform:** edge computing based on OpenStack

Our Way Forward in 2017
- Infrastructure, telco and research team up and build testbeds
- Application partners join the lab for dedicated test projects
- Integration and testing of latest edge computing applications
- Joint evaluation and promotion of results among partners

LEL

Carnegie Mellon University
T
CROWN CASTLE
Intel
Vodafone
NTT
Nokia

© 2017 M. Satyanarayanan
SEC Keynote Oct 12, 2017
Rest of This Talk

Does latency really matter?

Two “killer” use cases enabled by

• low end-to-end latency
• scalable bandwidth demand
Does Latency Really Matter?

"The Impact of Mobile Multimedia Applications on Data Center Consolidation"
Ha, K., Pillai, P., Lewis, G., Simanta, S., Clinch, S., Davies, N., Satyanarayanan, M.
Proceedings of IEEE International Conference on Cloud Engineering (IC2E), San Francisco, CA, March 2013

“Quantifying the Impact of Edge Computing on Mobile Applications”
Hu, W., Gao, Y., Ha, K., Wang, J., Amos, B., Pillai, P., Satyanarayanan, M.
Proceedings of ACM APSys 2016, Hong Kong, China, August 2016
Augmented Reality

E2E Response Time CDF

1. Send JPG image from device to cloud/cloudlet
2. Recognize landmark buildings using computer vision
3. Send labels & coordinates back to device

- Mobile-only
- Amazon East
- Amazon West
- Amazon EU
- Amazon Asia
- Cloudlet

4G LTE
T-Mobile for Cloud
In-lab Nokia eNodeB for Cloudlet
## Per-Operation Energy Use by Device

<table>
<thead>
<tr>
<th>Face Recognition</th>
<th>Augmented Reality</th>
</tr>
</thead>
</table>
| 12.4 J  
| 2.6 J  
| 4.4 J  
| 6.1 J  
| 9.2 J  | Mobile-only | 5.4 J  
|        | Cloudlet   | 0.6 J  
|        | Amazon East| 3.0 J  
|        | Amazon West| 4.3 J  
|        | Amazon EU  | 5.1 J  
|        | Amazon Asia| 7.9 J  |
What is the Killer Use Case?

“Towards Wearable Cognitive Assistance”
Ha, K., Chen, Z., Hu, W., Richter, W., Pillai, P., Satyanarayan, M.

“Early Implementation Experience with Wearable Cognitive Assistance Applications”
Chen, Z., Jiang, L., Hu, W., Ha, K., Amos, B., Pillai, P., Hauptmann, A., Satyanarayanan, M.

“An Empirical Study of Latency in an Emerging Class of Edge Computing Applications for Wearable Cognitive Assistance”
Chen, Z., Hu, W., Wang, J., Zhao, S., Amos, B., Wu, G., Ha, K., Elgazzar, K., Pillai, P., Klatzky, R., Siewiorek, D., Satyanarayanan, M.
Proceedings of SEC 2017, San Jose, CA, October 2017
A Unique Moment in Time

Convergence of Advances in 3 Independent Arenas

Cognitive Algorithms

Edge Computing

Wearable Hardware

This Research

Siri

Watson

DeepFace

Translator

Cloudlets

 ODG R7

Microsoft Hololens

Vuzix Wrap

Google Glass

Vuzix Wrap

Mobile Edge Computing Initiative
Industry Specification Group (ISG)
Bringing Compute and Storage to Base Stations

European Telecommunications Standards Institute

© 2017 M. Satyanarayanan
SEC Keynote Oct 12, 2017
Wearable Cognitive Assistance

A new modality of computing

Entirely new genre of applications

Wearable UI with wireless access to cloudlet

Real-time cognitive engines on cloudlet

• scene analysis
• object/person recognition
• speech recognition
• language translation
• planning, navigation
• question-answering technology
• voice synthesis
• real-time machine learning
• ...

Low latency response is crucial

Seamlessly integrated into inner loop of human cognition
Task-specific Assistance

Example: cooking

passive recipe display

versus active guidance

“Wait, the oil is not hot enough”
Inspiration: GPS Navigation Systems

Turn by turn guidance
  • Ability to detect and recover
  • Minimally distracting to user

Uses only one type of sensor: location from GPS

*Can we generalize this metaphor?*
Gabriel Architecture

(PaaS for Wearable Cognitive Assistance)

Device Comm → PubSub → UPnP → Context Inference → User Guidance VM → Control VM → Cognitive VMs

- Face recognition
- Object Recognition (MOPED)
- Object Recognition (STF)
- OCR
- Motion classifier
- Activity Inference
- Augmented Reality

Sensor flows → User assistance

Wireless connection

Wearable device

Video/Acc/GPS/… sensor streams

Sensor control

Cloudlet

© 2017 M. Satyanarayanan
Baby Steps: 2D Lego Assembly

Very first proof-of-concept (September 2014)

Deliberately simplified task to keep computer vision tractable

2D Lego Assembly  (YouTube video at http://youtu.be/uy17Hz5xvmY)
On Each Video Frame

(a) Input image
(b) Detected dark parts
(c) Detected board
(d) Board border
(e) Perspective corrected
(f) Edges detected
(g) Background subtracted
(h) Side parts added
(i) Unrotated
(j) Partitioned
(k) Synthesized

[Images of processing steps]

© 2017 M. Satyanarayanan SEC Keynote Oct 12, 2017
When Milliseconds Matter

Ping-pong assistant
(https://www.youtube.com/watch?v=_lp32sowyUA)
Assembling an IKEA Kit

IKEA kit assistant

(https://www.youtube.com/watch?v=qDPuvBWNIUs&index=5&list=PLmrZWttYffwHPy_4d61oDvQY_RBgS)
Many Monetizable Use Cases …

- Assembly instructions
- Medical training
- Industrial troubleshooting
- Correct Self-Instrumentation
- Strengthening willpower
AR Meets AI

Latency intolerance of Augmented Reality + Compute intensity of AI

October 9, 2016: CBS “60 Minutes” special on AI

- Short (90 seconds) video clip on Gabriel
- YouTube video at https://youtu.be/dNH_HF-C5KY
- Full 60 Minutes special (~30 minutes) at CBS web site: http://www.cbsnews.com/videos/artificial-intelligence
## Where Does the Time Go?

Attend Zhuo Chen’s talk tomorrow:

<table>
<thead>
<tr>
<th>Time</th>
<th>Session V – Performance and Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:30 – 12:45</td>
<td>Lunch</td>
</tr>
</tbody>
</table>
| 12:45 – 14:15| Edge Computing in the ePC - A Reality Check  
Zhuo Chen; Wenlu hu; Junjue Wang; Siyan Zhao; Brandon Amos; Guanhong Wu; Kiryong Ha; Khalid Elgazzar; Padmanabhan Pillai; Roberta Klatzky; Daniel Siewiorek; Mahadev Satyanarayanan |
|              | An Empirical Study of Latency in an Emerging Class of Edge Computing Applications  
Shanhe Yi; Zijiang Hao; Qingyang Zhang; Quan Zhang; Weisong Shi; Qun Li |
|              | LAVEA: Latency-aware Video Analytics on Edge Computing Platform  
Ilija Hadzic; Yoshihisa Abe; Hans Christian Woithe |
Edge Computing for Situational Awareness

“Edge Computing for Situational Awareness”
Satyanarayanan, M.
Proceedings of the 23rd IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN 2017), Osaka, Japan, June 2017

“Live Synthesis of Vehicle-Sourced Data Over 4G LTE”
Hu, W., Feng, Z., Chen, Z., Harkes, J., Pillai, P., Satyanarayanan, M.
Real-Time Data Overlaid on Maps
(periodic GPS location reports from participating vehicles)
Waze: Crowd-Sourced Human Annotations
(purchased by Google for ~$1B in 2013)
Best of Both Worlds?

*Detailed, Automated and Distraction-free*

**Computer vision** instead of just using GPS measurements

- 1+ video cameras on every vehicle
- video analytics to extract high-level information
- both driverless and drivered vehicles can contribute data

**Rich overlay of detailed information on map**

- road hazards  (potholes, dead animals, rocks, stalled cars, lane closures …)
- road conditions (fog, icy patches, deep snow, flooding, …)
- “street view” updates  (new store, old building torn down, …)
- … any other useful information that can be visually sensed/inferred

*Improve Situational Awareness*
Situational Awareness

“up-to-the-minute cognizance or awareness required to move about, operate equipment, or maintain a system”

highly mission-specific (broad interpretation of “mission”)

what matters is highly context-sensitive
Who Cares?

1. Local government
   - police chief, fire chief, road crews, …
   - where to direct scarce resources (salt trucks, fire trucks, patrol cars, …)
   - make better real-time decisions

2. Individual drivers
   - better anticipation of road conditions
   - better planning of travel
   - seamless integration with auto GPS

3. Driverless vehicles
   - acute need for up-to-date detailed map information
   - expensive to collect manually, why not crowd-source?
   - accurate maps allow proactive actions (rather than reactive)

4. Long-term planners
   - accurate and detailed information as free by-product
   - avoids expensive special-purpose data collection
Inspiration From the Past
(RAF Uxbridge, circa September 1940)

“Sensors” were radar stations with edge processing (human processors and communicators)

“Visualization” required you to sit at this vantage point in the room

Priceless in allocating scarce resources for survival (aircraft & pilots, just in time)
Fast Forward to 2017

End Users
- wide range of visualization devices
- handheld displays (tablets, smartphones)
- in-vehicle displays
- wall-sized displays...

Customizability
ability to download under tight access control mission-specific algorithms (e.g., object recognition)

Shared Information Repository

Mission-specific Software

Cloudlet-1
video analytics

Cloudlet-2
video analytics

Cloudlet-3
video analytics

Cloudlet-N
video analytics

associated cameras

associated cameras

associated cameras

Many Questions

1. Do we really need cloudlets?
2. Is computer vision up to the task?
3. How large a coverage area can we target?
4. How can we achieve scalability?
Do We Need On-board Cloudlets?

**Scarce wireless bandwidth**
- for the foreseeable future, vehicle connectivity will be 4G LTE
- already under severe pressure from customer demand
- limited and expensive spectrum, falling profit margins
- only small fraction can be spared for public service/safety

**Non-solution**
- rich highway infrastructure (e.g. roadside Wi-Fi)
- politically infeasible in the US
- may be feasible in other countries (e.g., Germany, Japan (?))
Consider small cell in Manhattan (2 block x 2 block)

- **roughly 400 vehicles under rush hour conditions**
  
  non-urban settings have lower vehicle density, but larger cells (evens out)

- **Netflix estimates 3 Mbps per SD video stream → 1.2 Gbps uplink demand**
  
  HD video is even worse (6.8 Mbps per video stream) → 2.7 Gbps uplink demand
  
  4K and future higher resolutions will be much worse
  
  higher resolution → improved accuracy, smaller features detectable

- **4G LTE uplink capacity is only ~500 Mbps**

- **5G will improve matters, but many other demands on wireless bandwidth**

**Shipping all video to cloud not scalable**

- 3-4 orders of magnitude lower demand with edge analytics in vehicle

- still true even if brief video clips or images accompany each report

- on-board cloudlet is crucial
Is Computer Vision Up to the Task?

Accuracy: challenging on diverse recognition tasks
- just within reach with deep neural networks
- very compute-intensive
  (need GPU or other specialized hardware)

Speed is important
- continuous processing of video for timeliness of reporting
- but less stringent than for V2V use cases (e.g., convoying, collision avoidance, …)
- recognition $\approx$ a few seconds at highway speeds (before object disappears)

Two examples
- deer detection (https://www.youtube.com/watch?v=_GrP42359z8)
- pothole detection (https://www.youtube.com/watch?v=U7_QAVbiF8U)
- only modest accuracy on classic metrics (e.g. ROC curve or precision/recall)
- acceptable accuracy for “detect before object disappears”
  “few seconds” $\rightarrow$ many hundreds of video frames, accuracy improves as object gets closer
- acceptable speed (7 FPS with high-end GPU on 3.4 GHz i7 using Faster R-CNN)

Lot more work ahead in terms of speed, accuracy, versatility, and reporting format

Basic premise ok
How Large a Coverage Area?

Ideal: entire planet

At least two reasons why this is unlikely

1. *end-to-end latency for near-real-time tracking of the real world*
   - both mean and variance matter: each network hop hurts

2. *national security, anti-terrorism, etc.*
   - nation-states unwilling to export fine-grain real-time street-level knowledge
   - static Street View of Google Maps is already causing angst

#2 is a showstopper
   - some deliberate degradation in timeliness or spatial resolution or both likely

More likely: *federation of autonomous zones*

- each zone retains full control over authentication and access control
- controlled cross-zone sharing possible

Zone size: city or county in US (~500 square miles) most likely
- coincides with local government boundary for organization control
Vehicle to Zone Cloudlet Interactions

Known Unknowns & Unknown Unknowns

1. Prioritized Data Stream of Observations
2. Current needs (priority information)
3. Updates to local cached knowledge
4. Drill-down or Ad hoc requests

Enables decentralized transmission control

red hot = recent, plentiful inputs
gray = old, sparse inputs

Heat Map Data Structure
cached everywhere
master copy at
zone cloudlet

Video Retention
- 3 GB per hour of HD video per camera
- single 4 TB disk → ~50 days of retention
- storage is cheap (~$100 for 4 TB disk)
Scalability Results

“Live Synthesis of Vehicle-Sourced Data Over 4G LTE”
Hu, W., Feng, Z., Chen, Z., Harkes, J., Pillai, P., Satyanarayanan, M.
Closing Thoughts
Navigating Edge Computing

1. “Let a thousand flowers boom”

Cloudlets will appear in many form factors and connectivities, with diverse levels of scale, management quality, and business models.

2. “One application, many cloudlets”

In spite of cloudlet diversity, an end-user application should see a single programming interface. Ideally, the same as in the cloud.

3. “The value chain begins with the end-user”

Without new applications that delight users and deliver long-term value to them, the business impact of cloudlets will be zero-sum.

4. “The edge is real, the cloud is abstract”

The new breed of latency-sensitive and bandwidth-hungry applications involve real-time processing of rich multi-sensor input streams using deep neural networks. These strongly suggest the need for application-specific hardware accelerators in cloudlets.
In Closing

*Edge Computing is transformative*

*It enables new applications*

*It is truly disruptive*

*It is here!*