Edge Computing

Vision and Challenges

Mahadev Satyanarayanan
School of Computer Science
Carnegie Mellon University

Classic Data Center

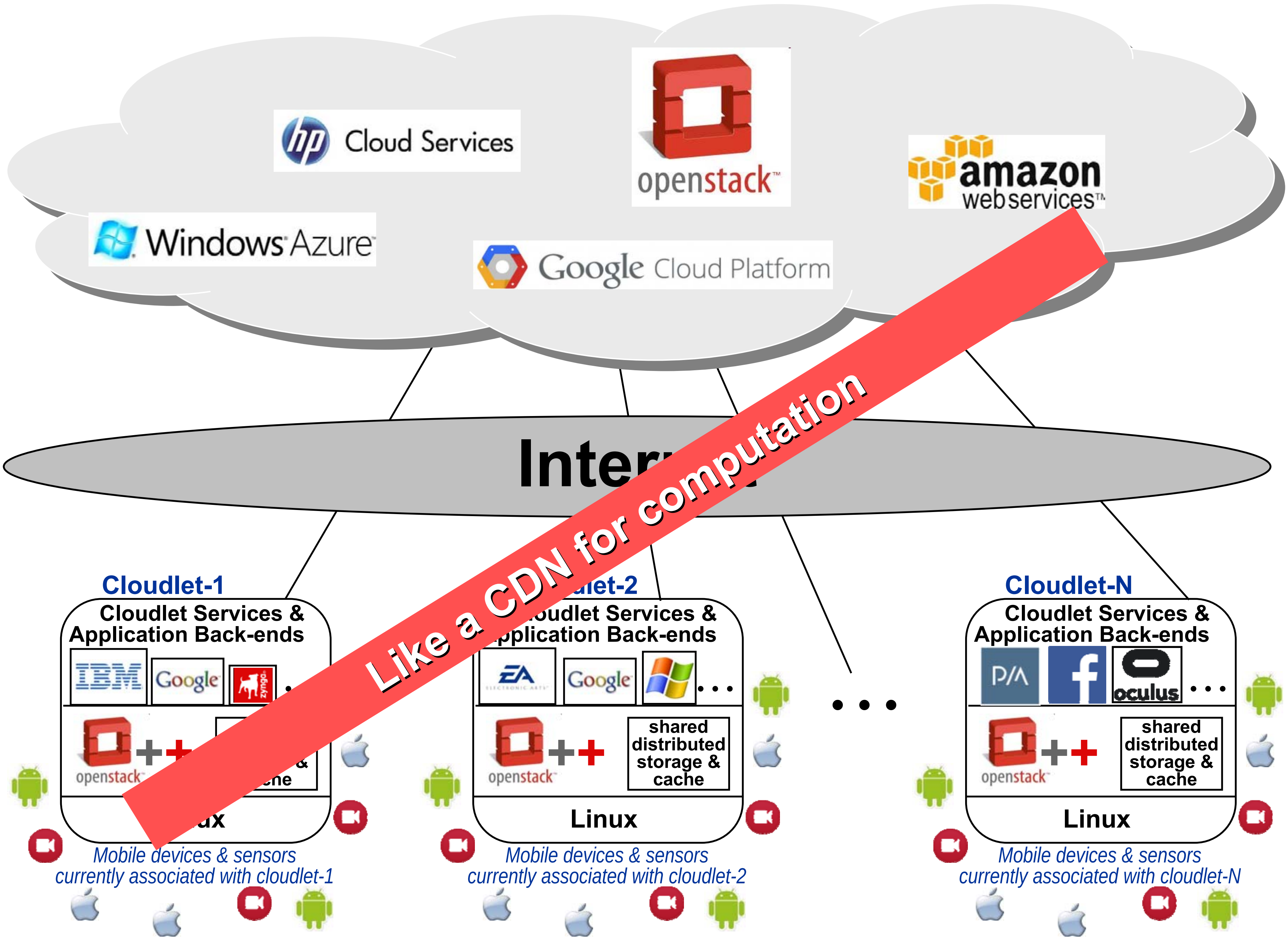
Strange Data Centers

Commercial Efforts Today

NOKIA
Flex base station and multicontroller

European Telecommunications Standards Institute
Mobile Edge Computing Initiative
Industry Specification Group (ISG)

Bringing Compute and Storage to Base Stations



Cloud Services
openstack
amazon webservices
Windows Azure
Google Cloud Platform
Internet
Like a CDN for computation
Cloudlet-1
Cloudlet Services & Application Back-ends
Mobile devices & sensors currently associated with cloudlet-1
Cloudlet-2
Cloudlet Services & Application Back-ends
shared distributed storage & cache
Linux
Mobile devices & sensors currently associated with cloudlet-2
Cloudlet-N
Cloudlet Services & Application Back-ends
shared distributed storage & cache
Linux
Mobile devices & sensors currently associated with cloudlet-N

What is a Cloudlet?

aka "micro data center", "mobile edge cloud", "fog node"

Small data center at the edge of the Internet

- **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**

Subordinate to the cloud
("second-class data center")

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

- **energy**
- **weight/size/heat**

Catalyst for new mobile applications

Value Proposition

1. **Highly responsive cloud services** — **Latency** (mean and tail)
   "New applications and microservices"

2. **Edge analytics in IoT** — **Bandwidth** (peak and average)
   "Scalable live video analytics"

3. **Exposure firewall in the IoT** — **Privacy**
   "Crossing the IoT Chasm"

4. **Mask disruption of cloud services** — **Availability**
   "Disconnected operation for cloud services"

How do we realize this value?

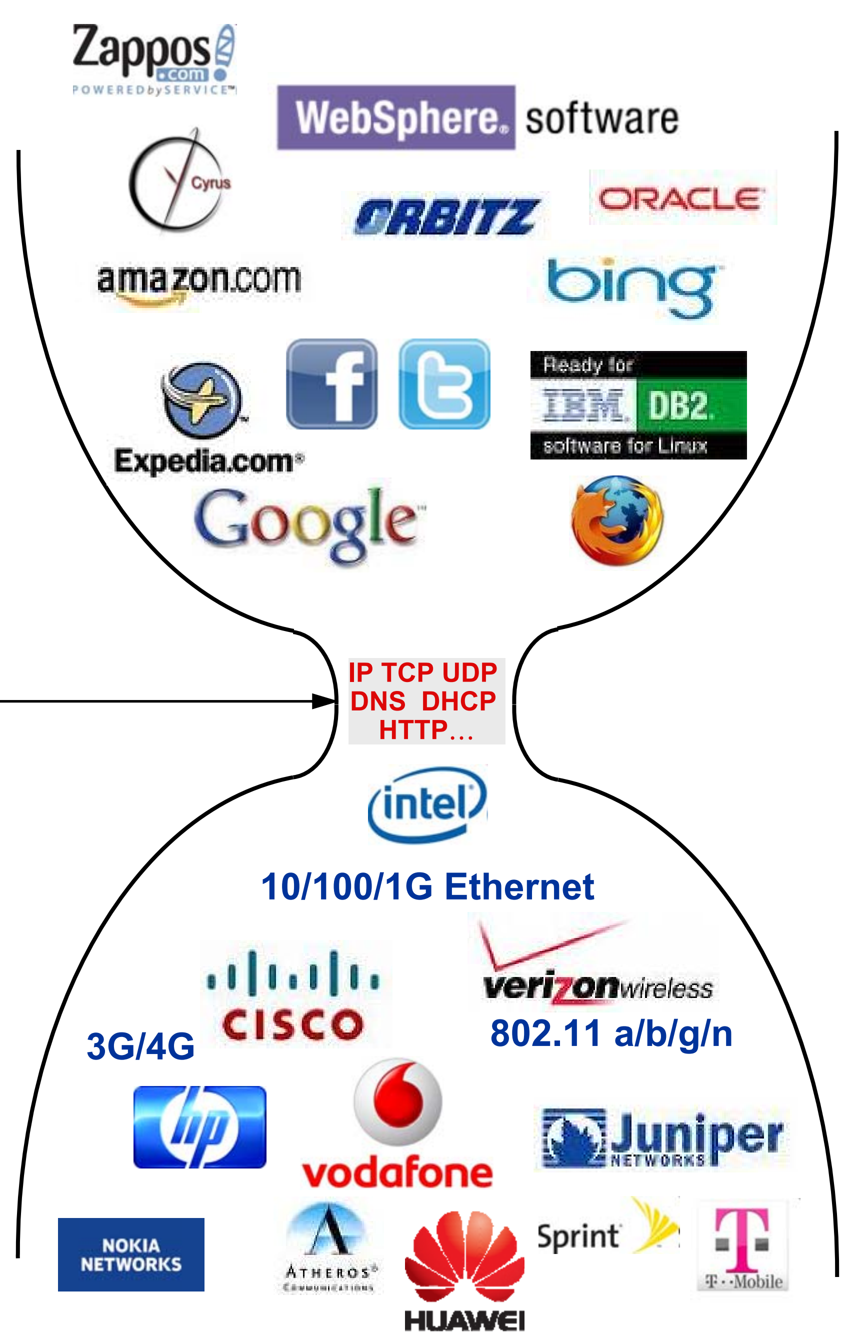
~$4T in 2013

G-20 Internet Economy

Barely a dozen protocols

~$1T in 2013

Total Market Cap



Internet Ecosystem

Getting There from Here

"We reject kings, presidents and voting. We believe in rough consensus and running code"

(attributed to Dave Clark of MIT, early Internet pioneer)

My own motto: *"Working Code Trumps All Hype"*

Focus on building and deploying real applications

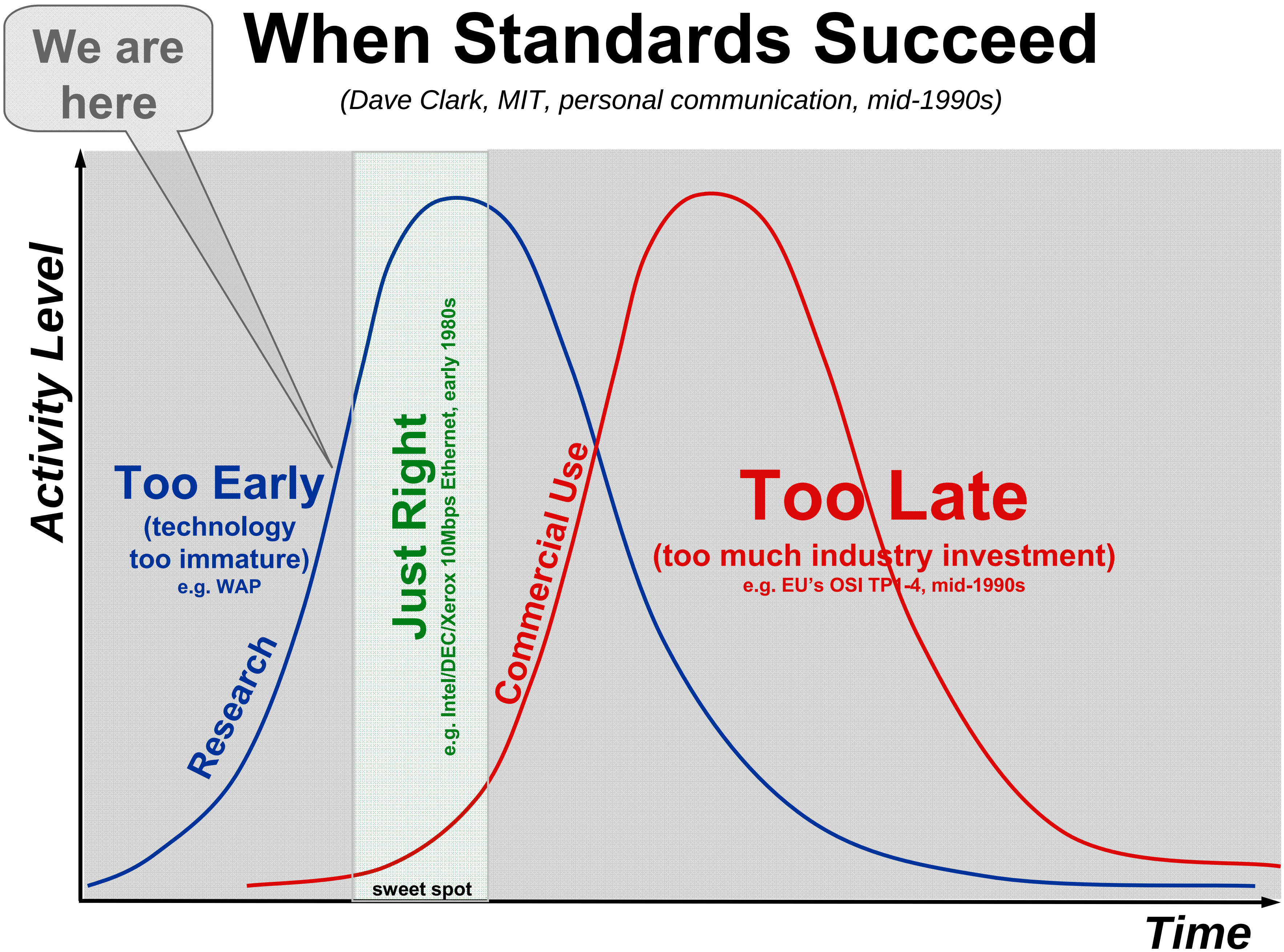
Work closely with companies

- learn real hands-on lessons from PoCs and pilots
- develop standards in the light of those lessons
- *premature standardization is worse than no standardization*

Deliver End-User Value

When Standards Succeed

(Dave Clark, MIT, personal communication, mid-1990s)



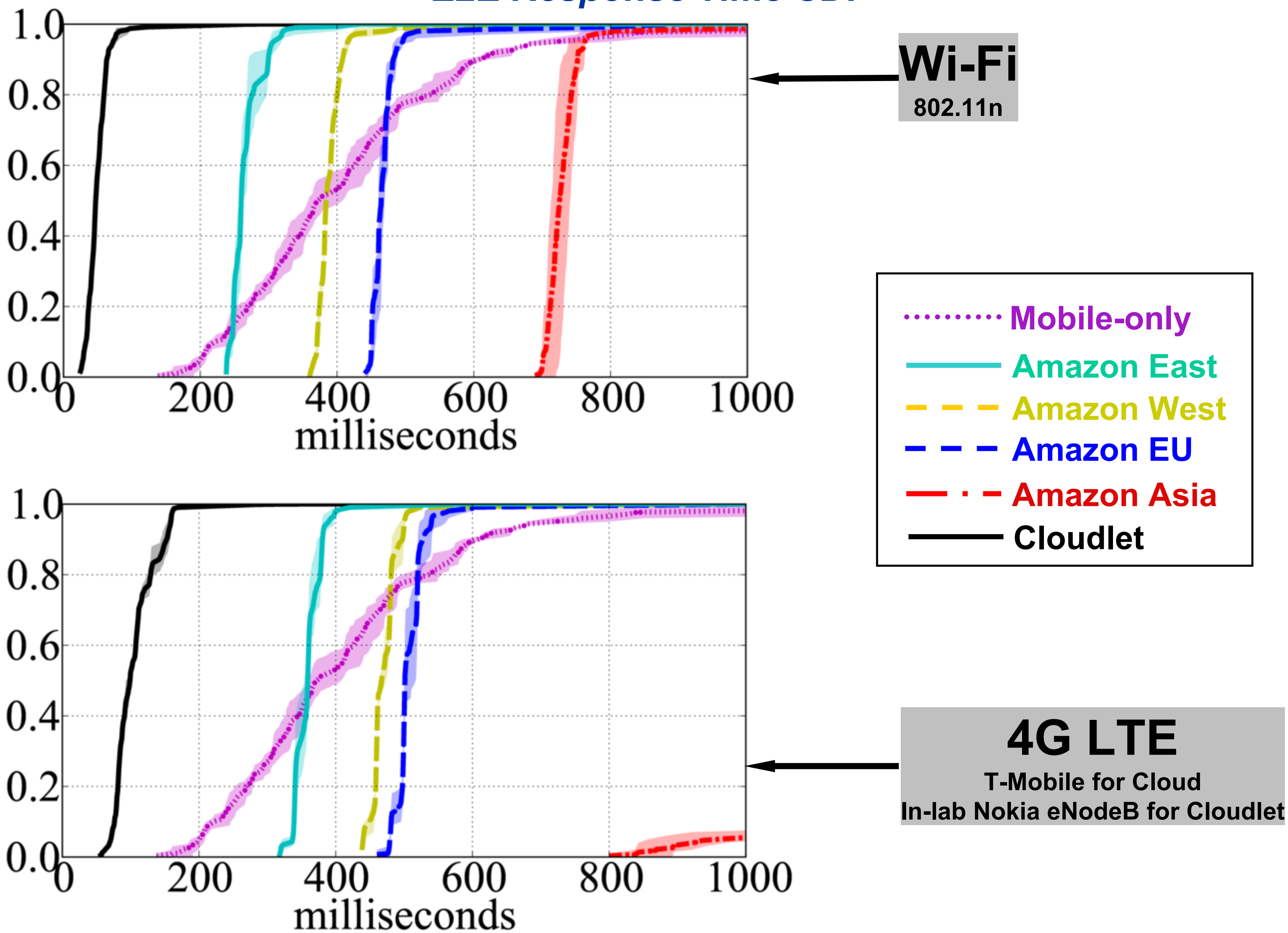
Does Latency Really Matter?

"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

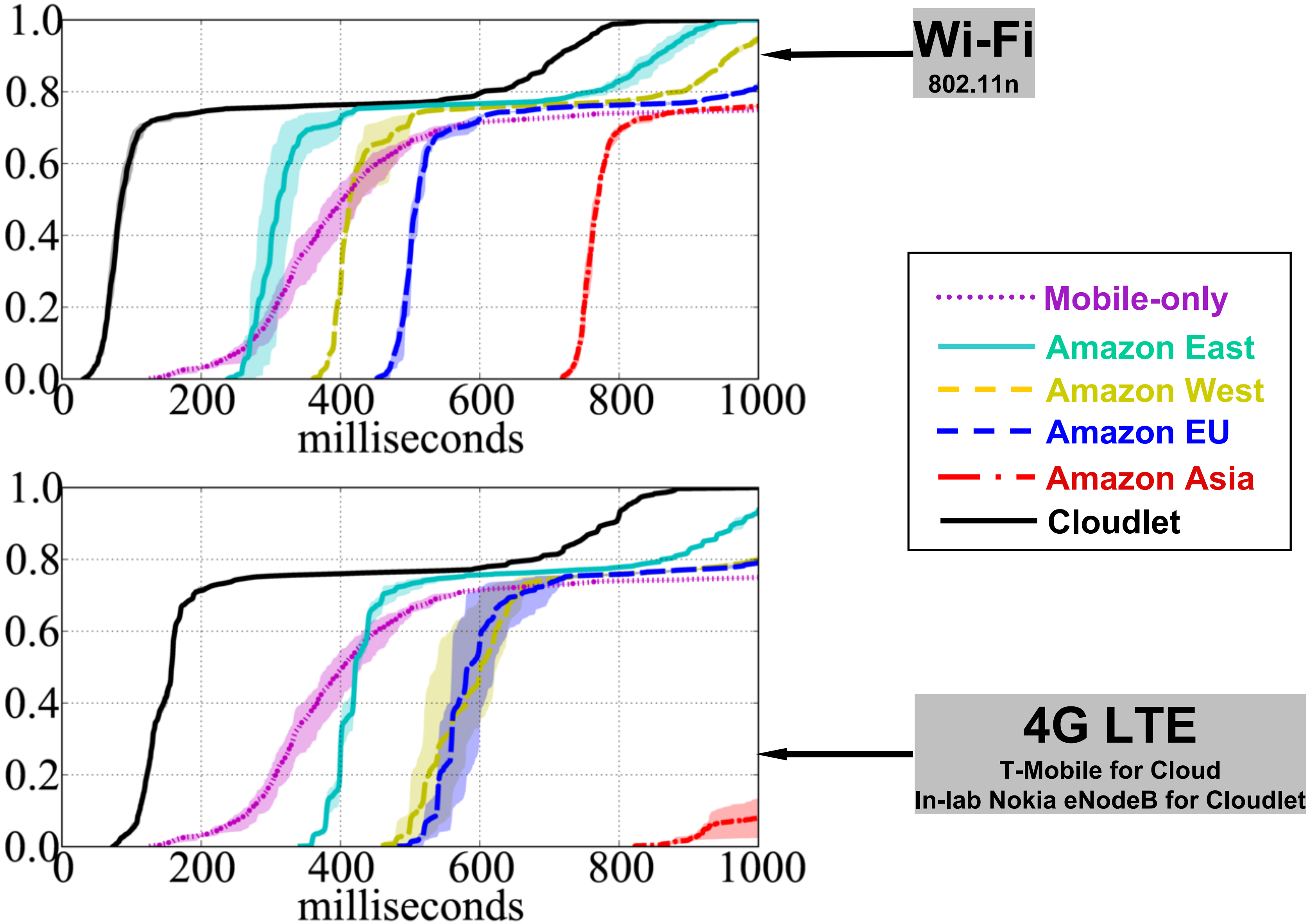
Augmented Reality

E2E Response Time CDF



Face Recognition

E2E Response Time CDF



Per-Operation Energy Use by Device

Face Recognition		Augmented Reality
12.4 J	Mobile-only	5.4 J
2.6 J	Cloudlet	0.6 J
4.4 J	Amazon East	3.0 J
6.1 J	Amazon West	4.3 J
9.2 J	Amazon EU	5.1 J
9.2 J	Amazon Asia	7.9 J

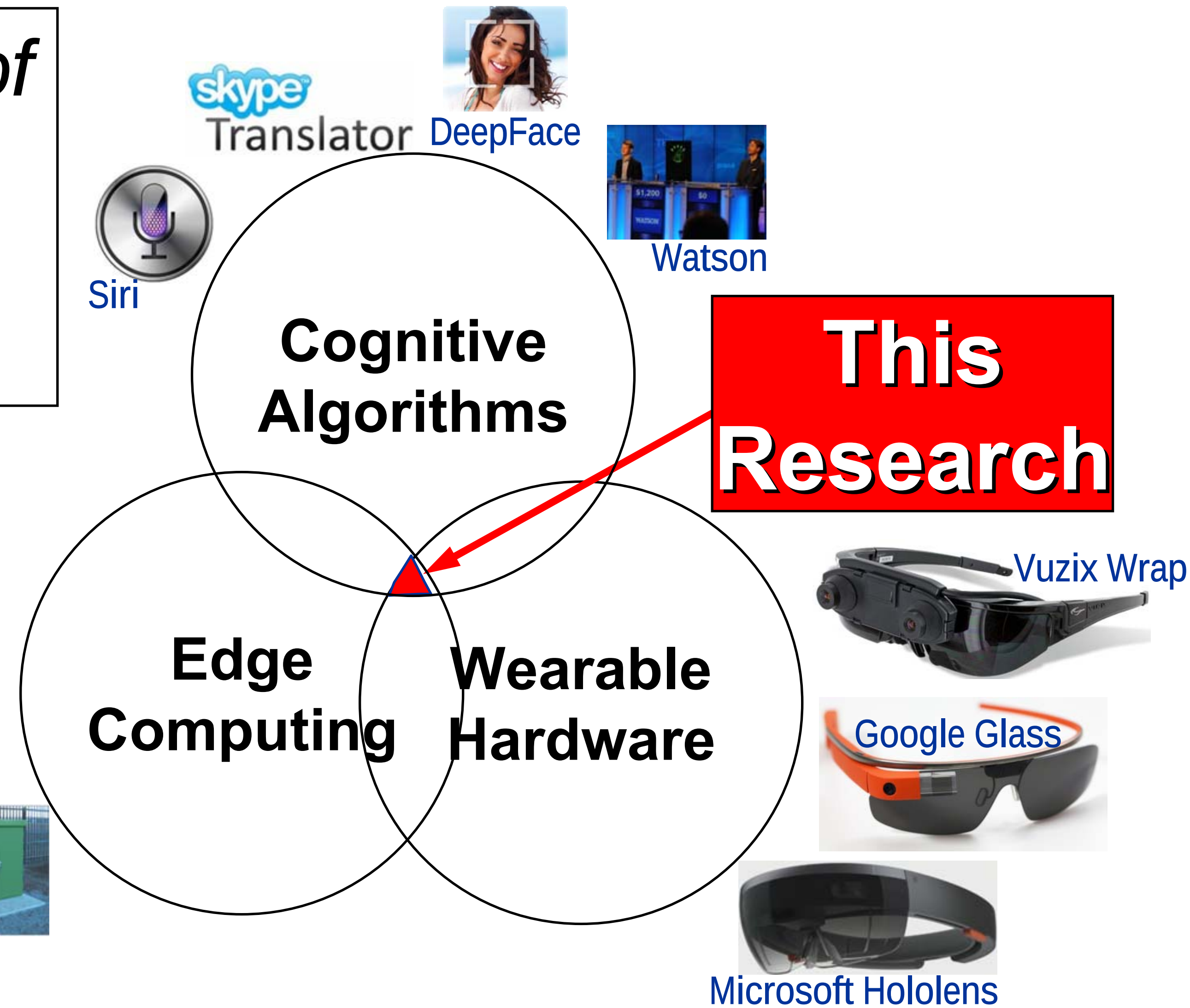
Latency: What is the Killer Use Case?

"Towards Wearable Cognitive Assistance"
Ha, K., Chen, Z., Hu, W., Richter, W., Pillai, P., Satyanarayanan, M.
Proceedings of the Twelfth International Conference on Mobile Systems, Applications, and Services (MobiSys 2014), Bretton Woods, NH, June 2014

"Early Implementation Experience with Wearable Cognitive Assistance Applications"
Chen, Z., Jiang, L., Hu, W., Ha, K., Amos, B., Pillai, P., Hauptmann, A., Satyanarayanan, M.
Proceedings of WearSys 2015, Florence, Italy, May 2015

A Unique Moment in Time

Convergence of Advances in 3 Independent Arenas



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

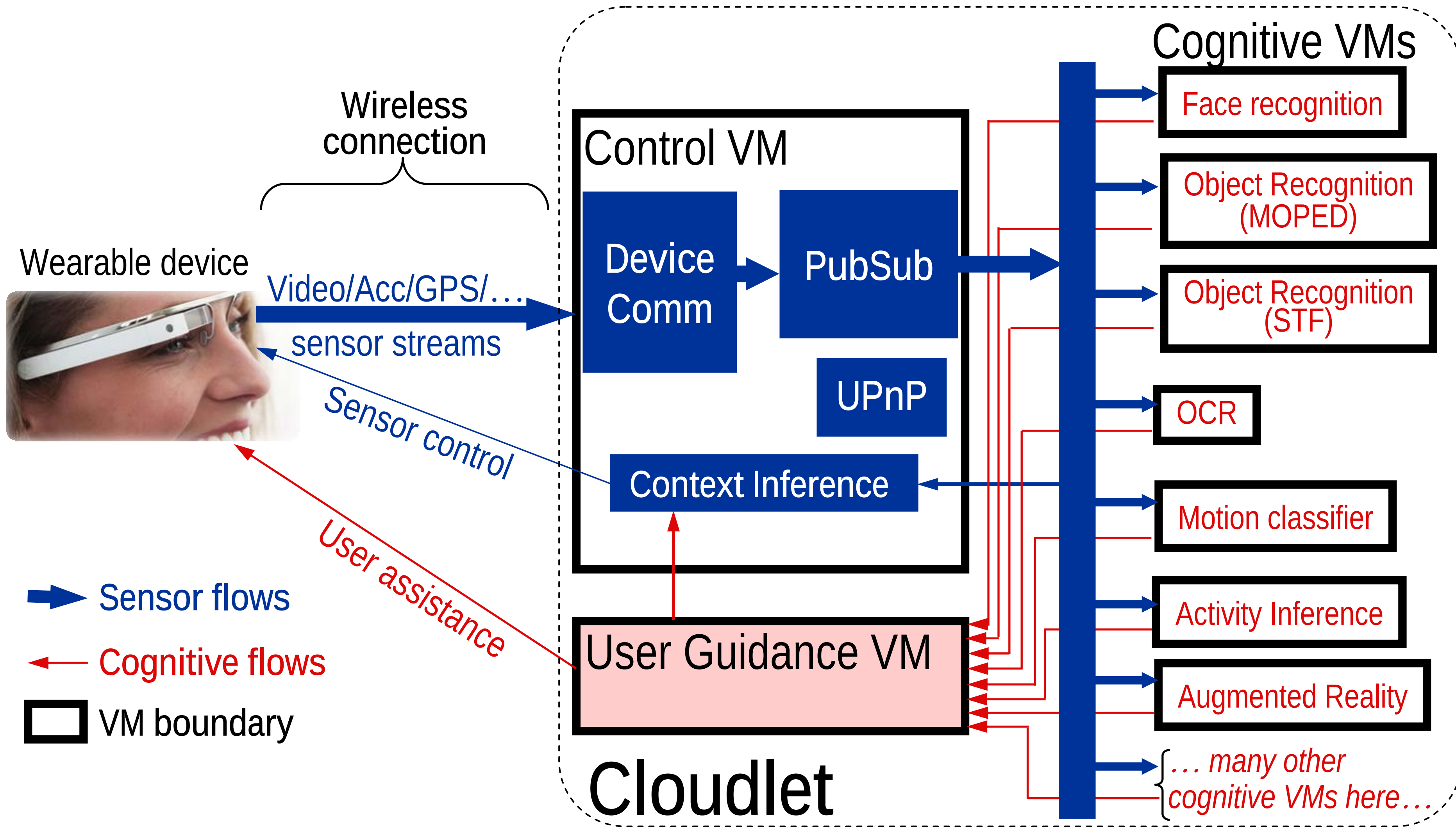
Turn by turn guidance

- **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



Baby Steps: 2D Lego Assembly

Very first proof-of-concept (September 2014)

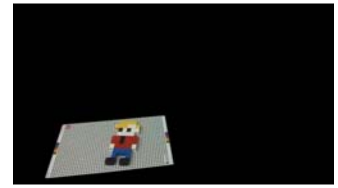
Deliberately simplified task to keep computer vision tractable

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

On Each Video Frame

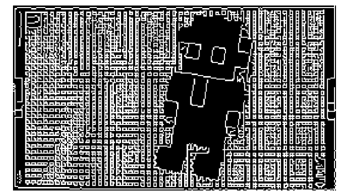

(a) Input image

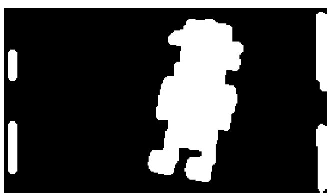

(b) Detected dark parts


(c) Detected board


(d) Board border


(e) Perspective corrected


(f) Edges detected

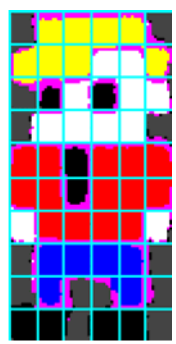

(g) Background subtracted


(h) Side parts added


(h) Lego detected


(i) Unrotated


(i) Color quantized


(j) Partitioned

[[0, 3, 3, 3, 3, 0],
[3, 3, 3, 1, 1, 3],
[0, 6, 1, 6, 1, 1],
[0, 1, 1, 1, 1, 0],
[4, 4, 6, 4, 4, 4],
[4, 4, 6, 4, 4, 4],
[1, 4, 4, 4, 4, 1],
[0, 5, 5, 5, 5, 0],
[0, 5, 0, 0, 5, 0],
[6, 6, 0, 6, 6, 0]]

(j) Matrix


(k) Synthesized

Example 2: Legacy Software

"Drawing by observation"

- corrective feedback for construction lines
- original version uses pen tablet and screen

Software developed at INRIA

"The Drawing Assistant: automated drawing guidance and feedback from photographs"

Iarussi, E., Bousseau, A. and Tsandilas, T.

In ACM Symposium on User Interface Software and Technology (UIST), 2013.

The Drawing Assistant: Automated Drawing Guidance and Feedback from Photographs

Emmanuel Iarussi
REVES / Inria Sophia Antipolis

Adrien Bousseau
REVES / Inria Sophia Antipolis

Theophanis Tsandilas
Inria - Univ Paris-Sud & CNRS (LRI)

(a) Interaction setup (b) Model and extracted guides (c) User construction lines and drawing

Figure 1. Our drawing assistant provides guidance and feedback over a model photograph that the user reproduces on a virtual canvas (a). We use computer vision algorithms to extract visual guides that enhance the geometric structures in the image (b). In this example, the user first sketched the block-in construction lines (c, blue) before drawing the regions and adding details. This guidance helps users produce more accurate drawings.

Our goal

- use Google Glass to untether this application
- allow drawing using any medium in the real world
  (paper, whiteboard, oil paint and brush on canvas, etc.)

Visual Feedback

Drawing assistant

(https://www.youtube.com/watch?v=nuQpPtVJC6o)

Example 3: When Milliseconds Matter

Ping-pong assistant

(https://www.youtube.com/watch?v=_lp32sowyUA)

Many Monetizable Use Cases …

Assembly instructions

Industrial troubleshooting

Medical training

Correct Self-Instrumentation

Strengthening willpower

AR Meets AI

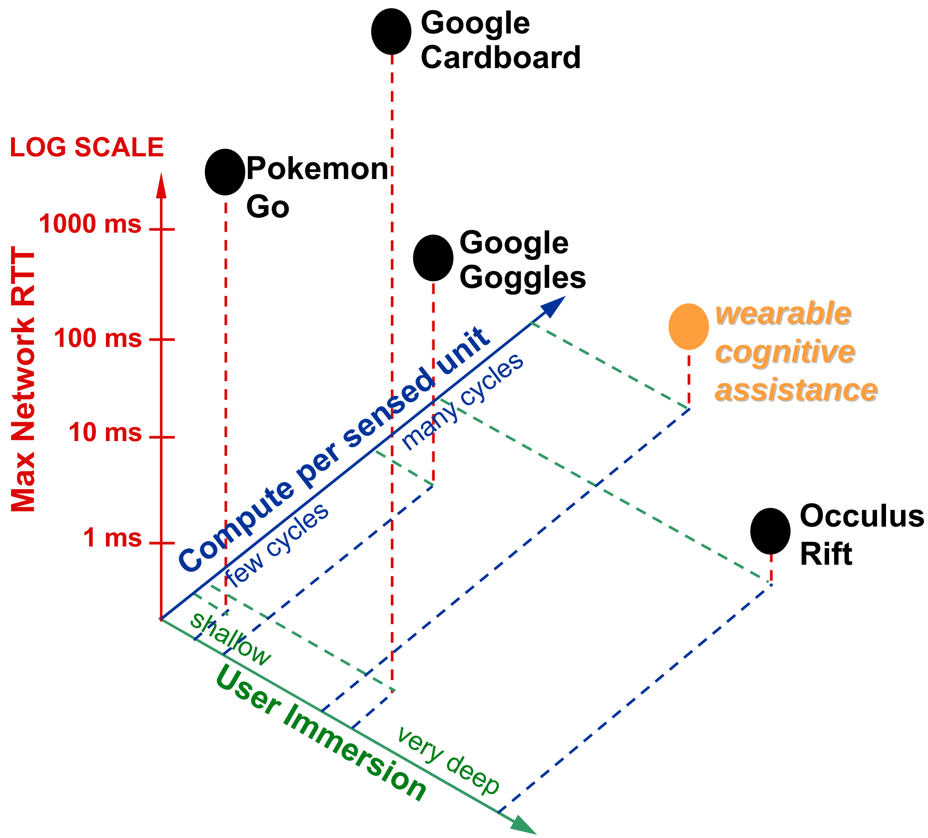
Low latency of AR + Compute intensity of AI

Has the “look and feel” of AR, but the functionality 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

Augmented Reality Taxonomy



Pokemon Go

- Shallow immersion
- Almost zero computation
- Mostly on smartphone
- Use of GPS and remote game info

Google Cardboard

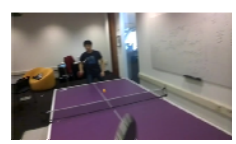
- Deep immersion
- Almost zero computation
- Entirely on smartphone
- No offload, so infinite RTT OK

Google Goggles

- Shallow
- Light computation
- Google cloud
- ~100-1000 ms RTT OK

Oculus Rift

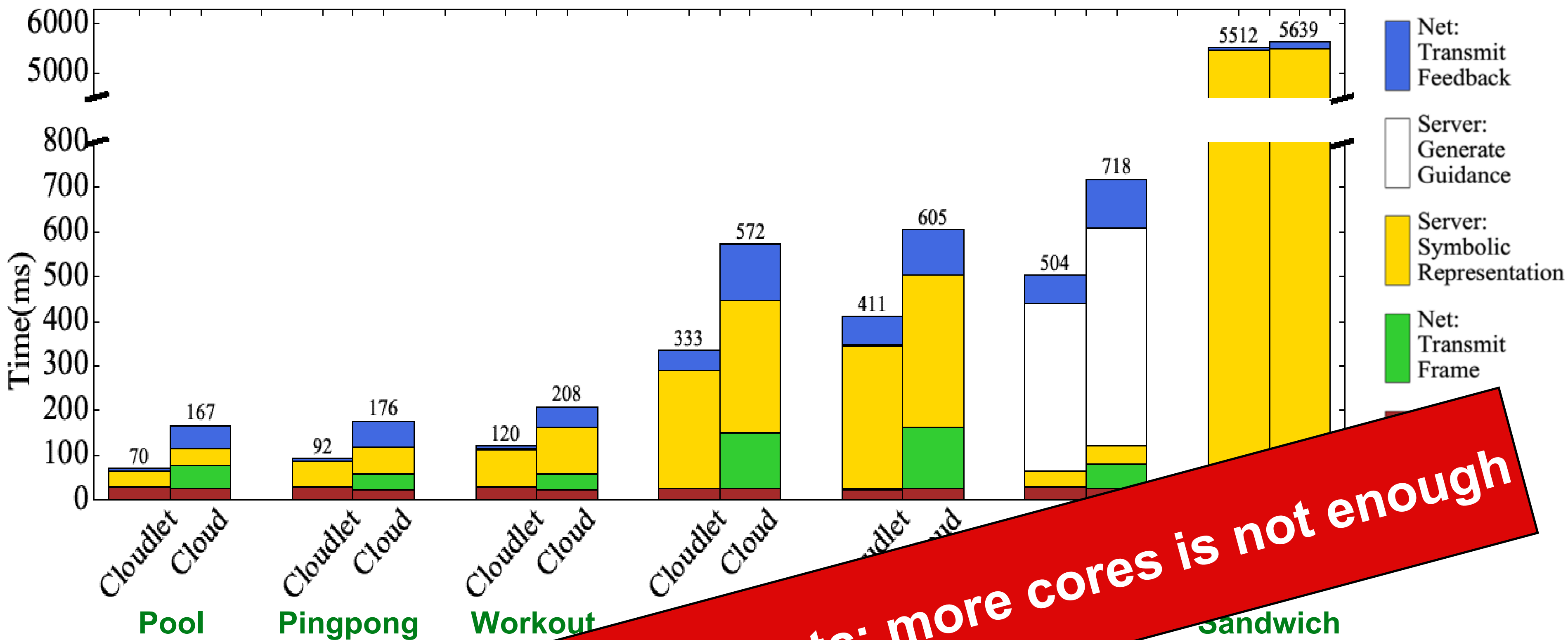
- Very deep immersion
- Intense computation
- Dedicated PC
- ~1 ms (tethered only)



Wearable Cognitive Assistance

- Medium immersion
- Intense computation
- Cloudlet
- ~10-30 ms

Where Does Time Go?



Cloudlets need GPUs, ASICs, etc; more cores is not enough

Network ti... ...ween cloudlet & cloud

...s similar on cloudlet and cloud

S...wich is huge outlier: deep neural network (DNN) classifier w/o GPU

Bandwidth: Edge Analytics for IoT Video

"Scalable Crowd-Sourcing of Video from Mobile Devices"

Simoens, P., Xiao, Y., Pillai, P., Chen, Z., Ha, K., Satyanarayanan, M.

Proceedings of the Eleventh International Conference on Mobile Computing Systems, Applications and Services (MobiSys 2013), Taipei, Taiwan, June 2013

"Edge Analytics in the Internet of Things"

Satyanarayanan, M., Simoens, P., Xiao, Y., Pillai, P., Chen, Z., Ha, K., Hu, W., Amos, B.
IEEE Pervasive Computing, Volume 14, Number 2, April-June 2015

"A Scalable and Privacy-Aware IoT Service for Live Video Analytics"

Wang, J., Amos,B., Das, A., Pillai, P,, Sadeh, N., Satyanarayanan, M.
Proceedings of the 2017 ACM Multimedia Systems Conference, Taipei, Taiwan, June 2017
.,

Video Cameras are Everywhere



"One surveillance camera for every 11 people in Britain, says CCTV survey."

Daily Telegraph (July 10, 2013)

"It will soon be possible to find a camera on every human body, in every room, on every street, and in every vehicle."

NSF Workshop on "Future Directions in Wireless Networking," Arlington VA, November 4-5, 2013

Why Video is So Powerful

Unique attributes for sensing

- **non-invasive** (as opposed to embedded sensors)
- **very high resolution**
- **large coverage area**
- **flexibility after installation** (new video analytic algorithms)
- **direct comprehensibility by humans** (skip the ML-based inference step)

Suggests *"Video as an IoT Service"*

- **provide live video feed to third-party video analytics**
- ***pre-processed to strip privacy-sensitive pixels***
- **a unique, monetizable resource**

Valuable Extractable Knowledge

Missing Child

Icy Sidewalk

Spilled Liquid

Ignored Display

Long Line

Scary Bandwidth Demand

Video analytics is typically done in the cloud

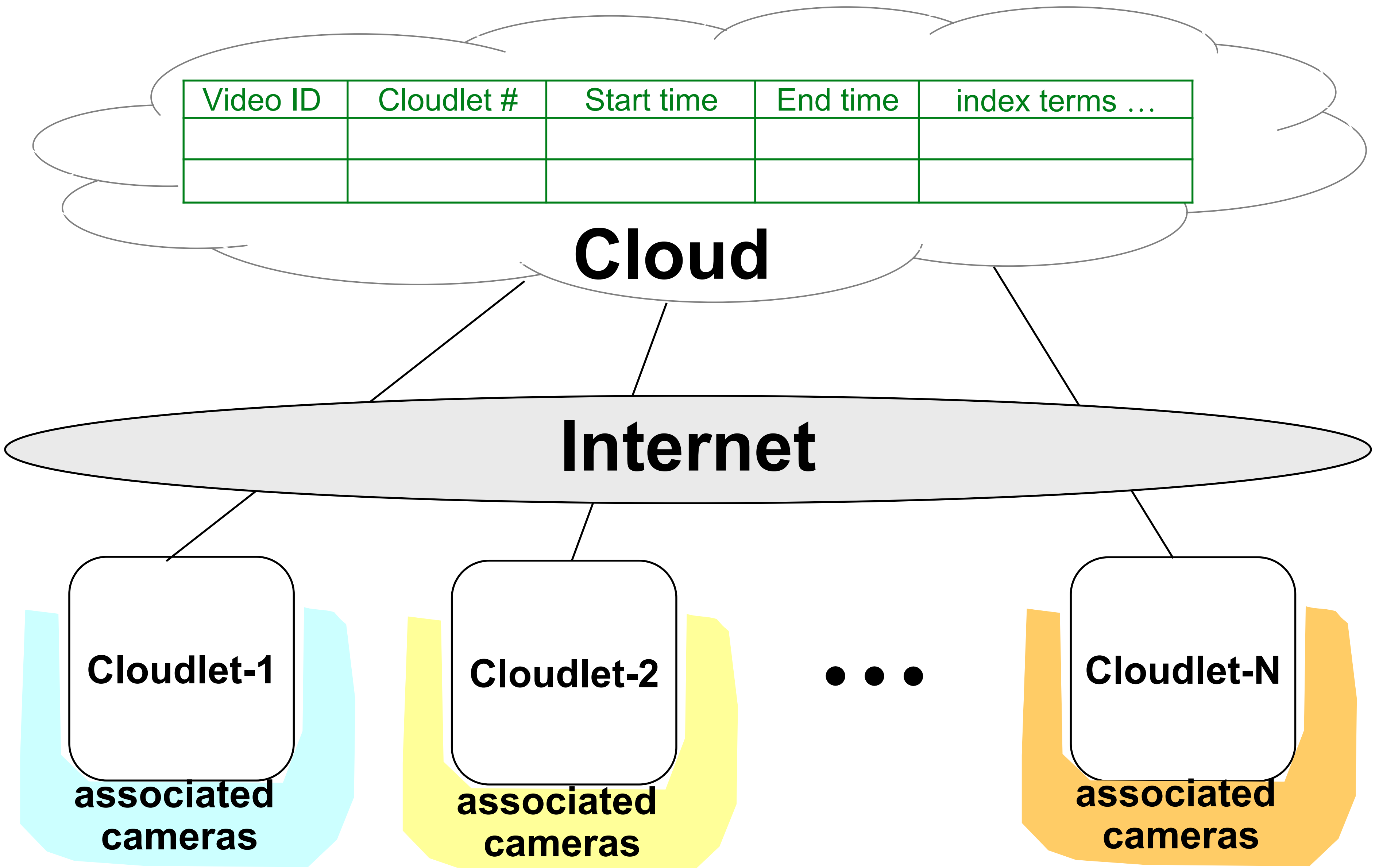
Shipping all the video to cloud is not scalable

- **Netflix estimate: 3 GB/hr of HD video → 6.8 Mbps per stream**
- **typical ingress MAN is 100 Gbps → ~15,000 HD video streams**
- **even upgrade to 1 Tbps only supports ~150,000 video streams**
  London is estimated to have 500,000 surveillance cameras today
- **1 million cameras would require ~ 7 Tbps**
- ***this is continuous demand: no "off-peak" period***

Future demand even higher: higher resolution video (e.g., 4K and beyond)

Only solution: Edge Computing

Edge Analytics for Video



Key Challenge: Privacy

Secondary challenge: scalability

Big concern and potential showstopper

- **face recognition can be part of the solution**
- **e.g. "don't ever record John's face; blur it before recording"**
  but no need to blur Donald Trump's face ever
  develop principles for responsible public use

"Denaturing" **= policy-guided reduction of fidelity of IoT data**

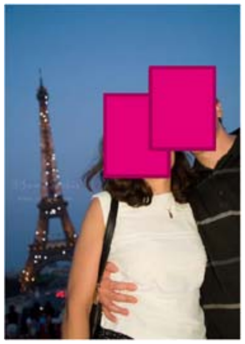
- **makes data safe for public release**
- **user-specific denaturing of streamed video is possible**
- **by definition content-based, but can also leverage meta-data**
  (e.g. timestamp, location, etc.)

Classic separation of *policy* and *mechanism*

- **our focus is on scalable mechanism**
- **informed by likely range of desirable policies**

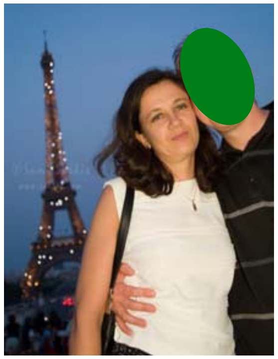
Examples of Denaturing

Blur all faces



+ removal of
location cue

Selective face blurring



blank video →
perfect privacy but zero value

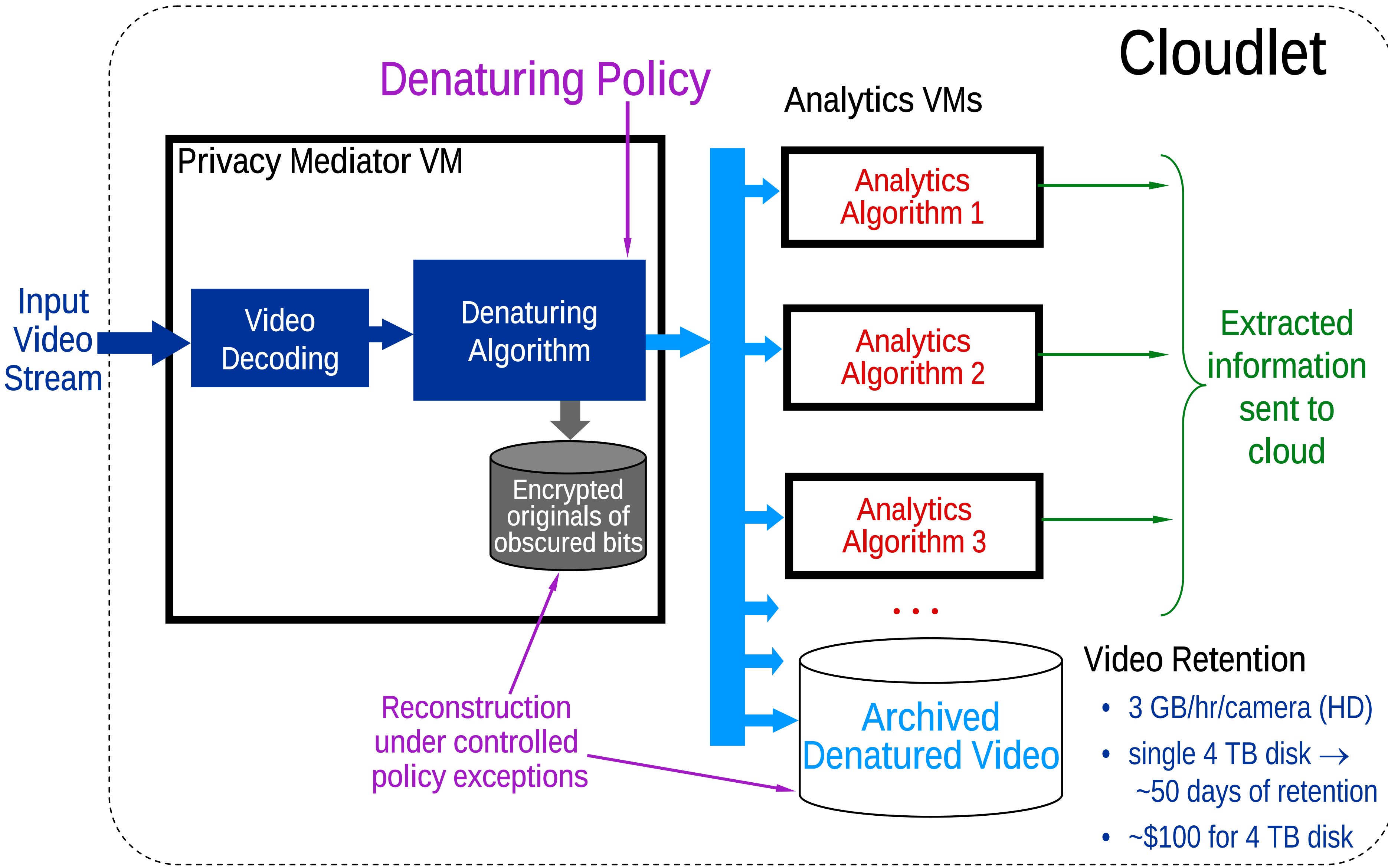
original video →
highest value but least privacy

Blur video segments based
on activity detection
(exactly how "blur" is done remains to be defined)

Shaking hands OK

Blur intimate scene

On Each Cloudlet



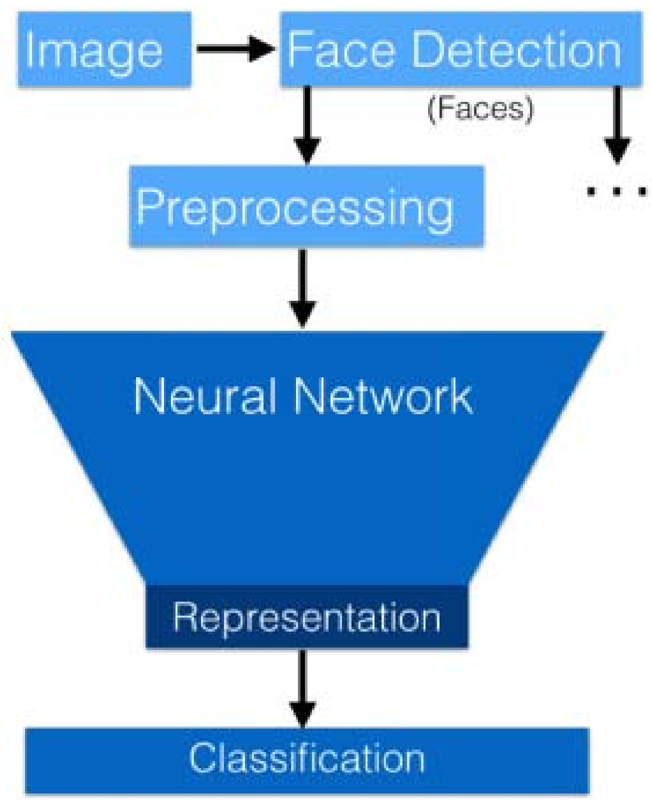
OpenFace

Inspired by FaceNet (CVPR 2015)

Trained with 500K images from public dataset
(CASIA-WebFace + FaceScrub)

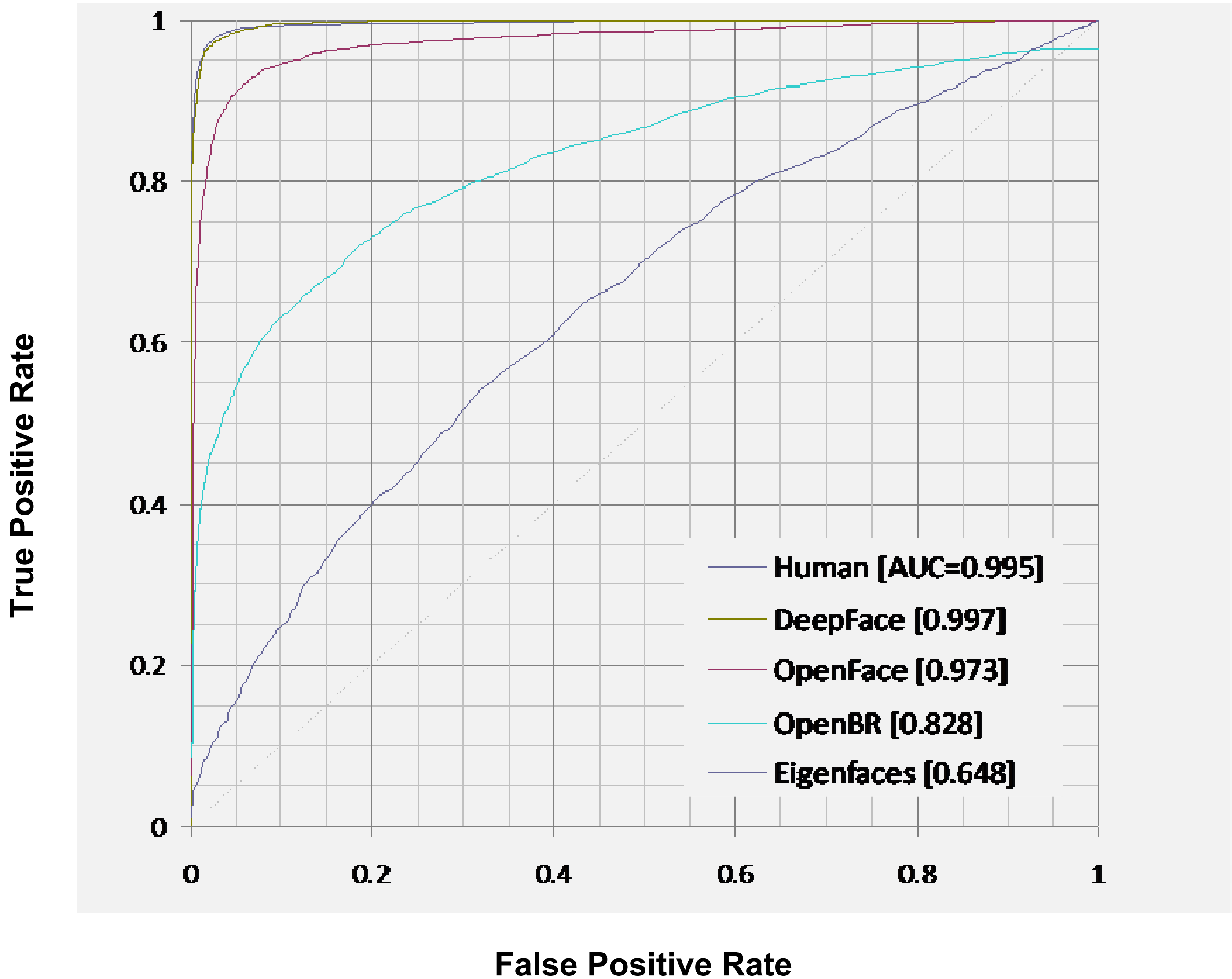
DNN + SVM Approach

- **DNN to extract facial features**
- **SVM to classify faces**



Accuracy on LFW benchmark

Predict whether pairs of images are of the same person



Simple Pipeline is Too Slow

figures in parens are
standard deviations from 3 runs

OpenFace	60 (28) ms per face *(easy per-face parallelism)*
Face detection (Dlib)	127 (1) ms per frame
Face tracking	11 (3) ms per face *(easy per-face parallelism)*
Perceptual hashing	0.3 (0.1) ms per BB

Intel 4-core i7-4790
with HyperThreading, no GPU
(high end desktop)

GPU only speeds recognition
(not detection)

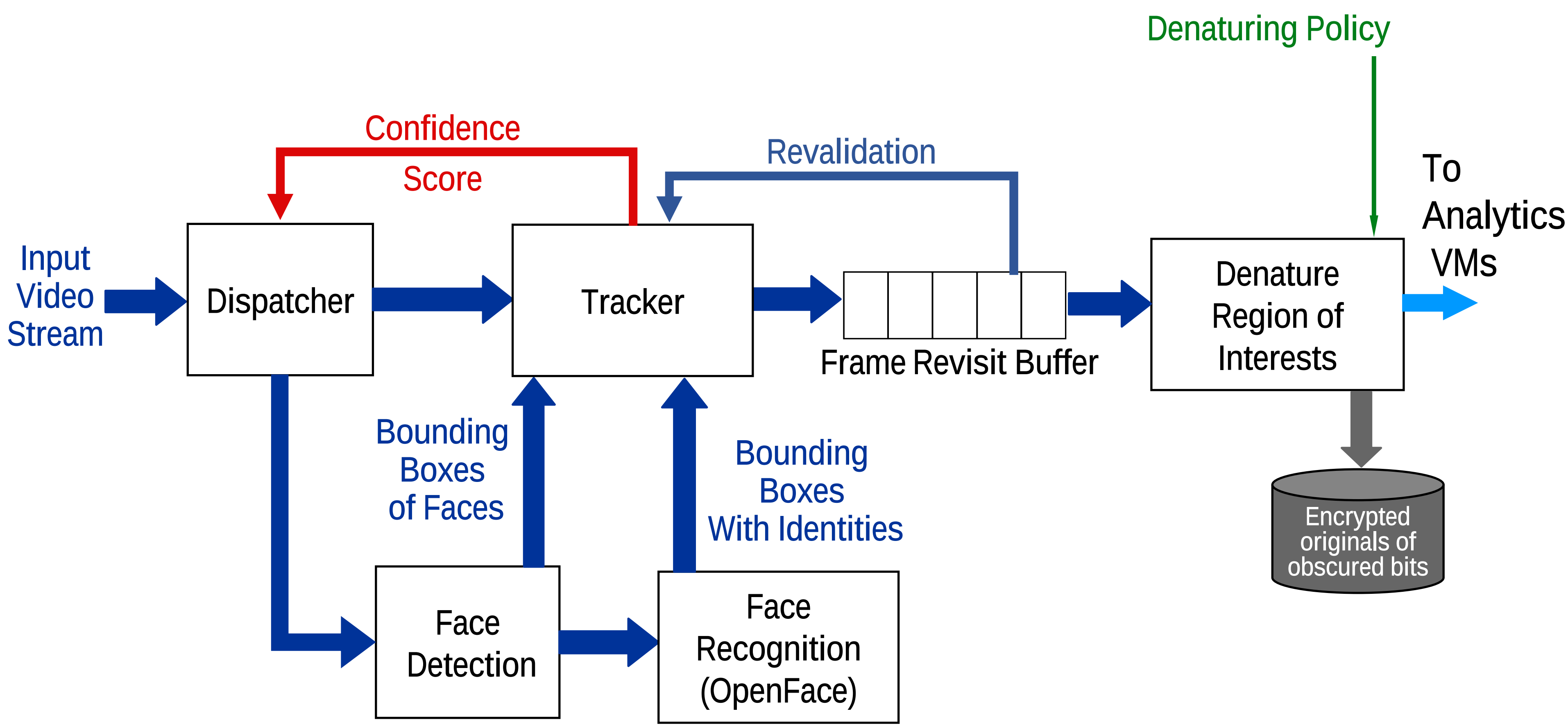
30 fps → ~33 ms to find all faces, recognize each, then denature per policy

Just the first two add up to more than 180 ms!

Solution strategy

- *faces don't move dramatically across two consecutive frames*
- at most small translation of pixels (even athletic movement)
- use *face tracking* to lower processing cost after recognition

Combine Face Detection & Tracking



Force detection on

- **low confidence in tracking**
- **every N frames**

Speed on same hardware is ~31 fps

Many Details Skipped

see MMSys 2017 paper

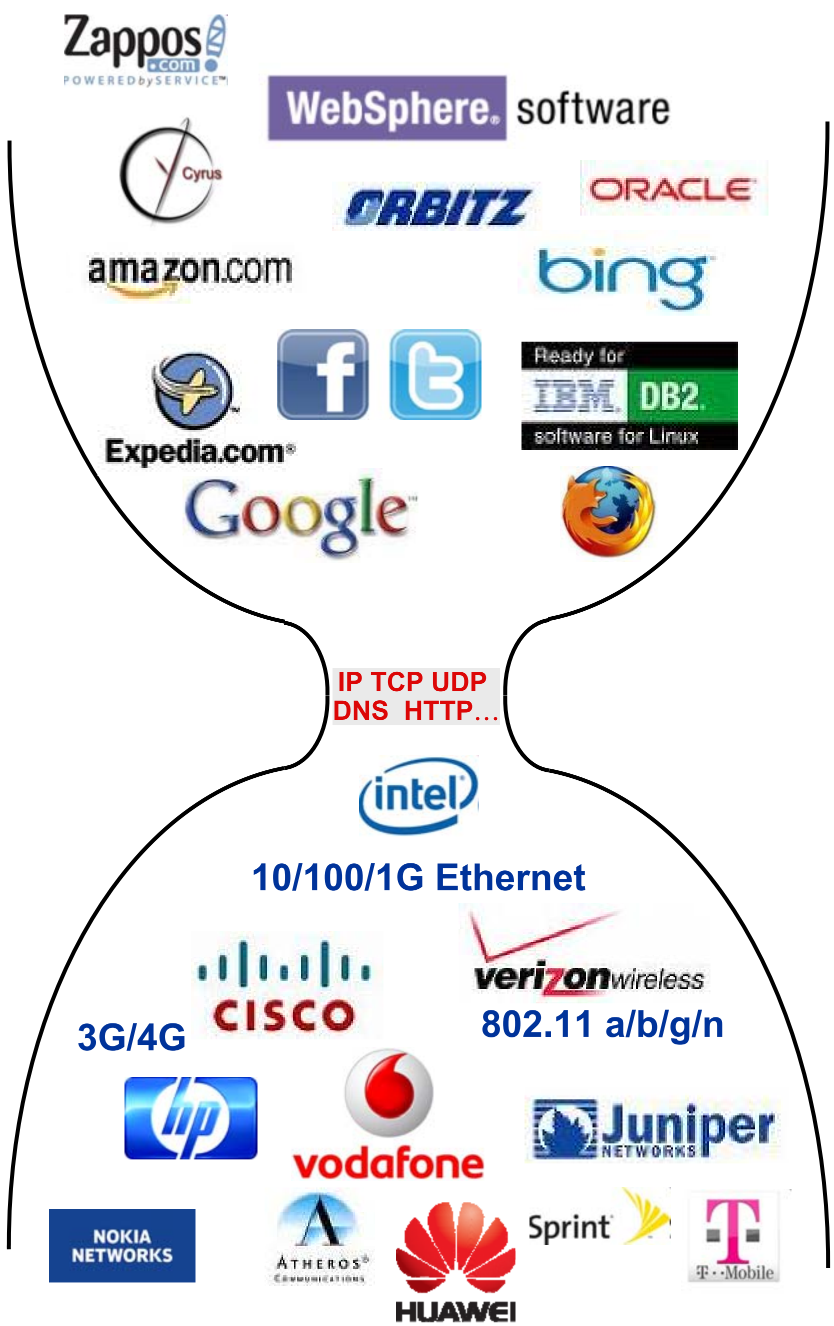
- Optimizations to Reduce Privacy Leaks
- Controlled Reversal of Denaturing
- IoT Service Deployment at Enterprise Scale
- Design Choices for Cloudlets

Enabling This New World

"An Open Ecosystem for Mobile-Cloud Convergence"
Satyanarayanan, M., Schuster, R., Ebling, M., Fettweis, G., Flinck, J., Joshi, K., Sabnani, K.
IEEE Communications Magazine, Volume 53, Number 3, March 2015

RECAP

~$4T in 2013
G-20 Internet Economy



~$1T in 2013
Total Market Cap

Open Internet Ecosystem

OpenStack++

Open Stack ≈ EC2-like cloud services and REST API

- **Apache v2 open source license**
- **widely used in industry**
  (HP, Dell, IBM, Intel, Oracle, NetApp, CloudBase, CloudByte, CloudScaling, Piston Cloud, …)
- **APIs for commonly used cloud services and management**
  (identity, compute, image, object storage, networking, block storage, …)

Extensions for cloudlets (Kiryong Ha, PhD thesis, December 2016)

1. **Cloudlet discovery**
2. **Rapid cloudlet provisioning (dynamic VM synthesis) (MobiSys 2013)**
3. **Adaptive VM handoff across cloudlets (SEC 2017)**

In Closing

Eternal battle between forces of *centralization* and *dispersion*

Has produced epochal, decade-long shifts in the nature of computing

- **batch processing** (centralized compute, centralized access)
- **timesharing** (centralized compute, dispersed access)
- **personal computing** (dispersed compute, dispersed access)
- **web-based distributed systems** (dispersed compute, dispersed access)
- **cloud computing** (centralized compute, dispersed access)

Edge computing is the latest chapter of this long-running story

Edge Computing enables an exciting new world