Towards Wearable Cognitive Assistance

Presenter: Wenqi Xu

Cognitive Decline

- 20 million Americans are affected by cognitive decline

  survivors of stroke; mild cognitive impairment; Alzheimer disease

- Cognitive decline can manifest itself in many ways

  inability to recognize people, locations and objects

- One month delay in nursing home admissions in the US could save over $1 billion


Faces

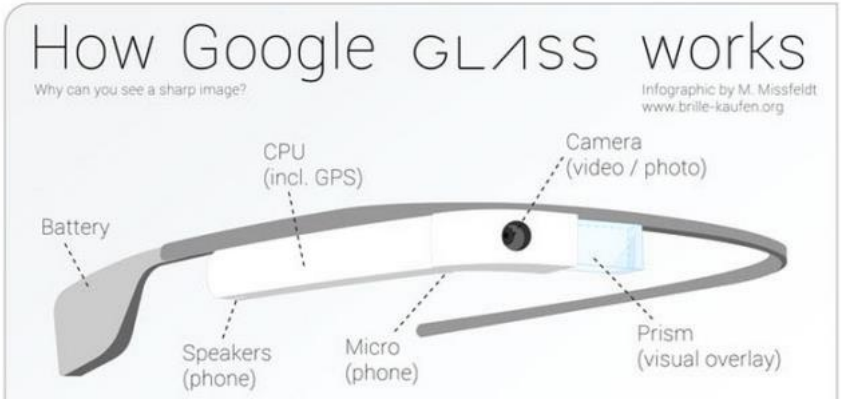

Text


Daily Routine

Can Wearable Technology Help?

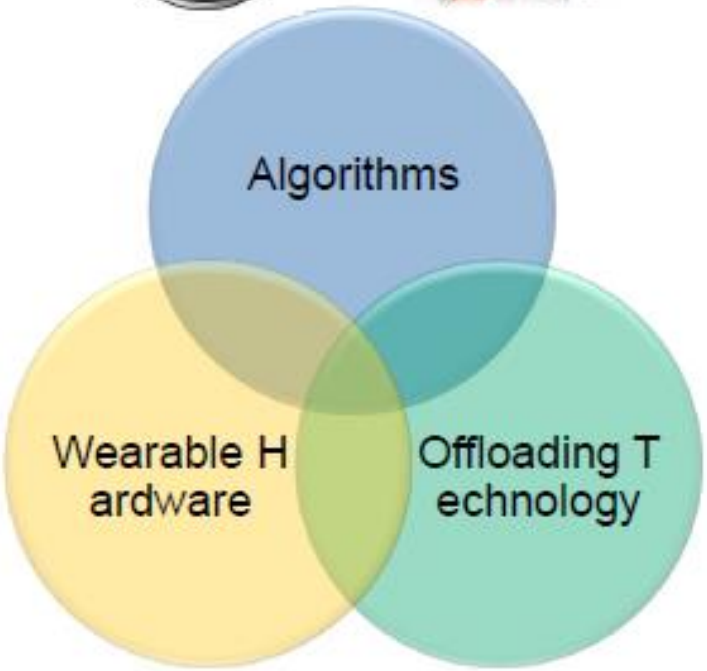
- Wearable devices such as *Google Glass* offer a glimmer of hope to users in cognitive decline.
- Continuously capture, interpret, and give guidance
- System Architecture:

  All sensors see what you see and hear what you hear;
  Processing the sensor input in real time on a cloud;
  Getting result faster than a person can think;
  Give guidance to users saying something you could do



Hypothetical Scenario of Cognitive Assistance

Why Today? Advances in 3 Areas



Challenges

1. Crisp Interactive Response

2. Graceful Degradation of Services

3. Coarse-grain Parallelism

Challenge 1: Crisp Interactive Response

- Humans are amazing fast, accurate and rubust

face detection under hostile condition	*< 700 ms*
face recognition	*370 – 620 ms*
is this sound from a human	*4 ms*
VR head tracking	*< 16 ms*

Goal: Latency of infrastructure = tens of millisecond

Conquering Latency

- Choice 1: Standalone apps 

Offloading saves latency and energy

Offloading vs. Standalone (OCR)

Metric	Standalone	With Offload
Per-image speed (s)	10.49	1.28
Per-image energy(J)	12.84	1.14

- Choice 2: offload to cloud 

  RTT is too long

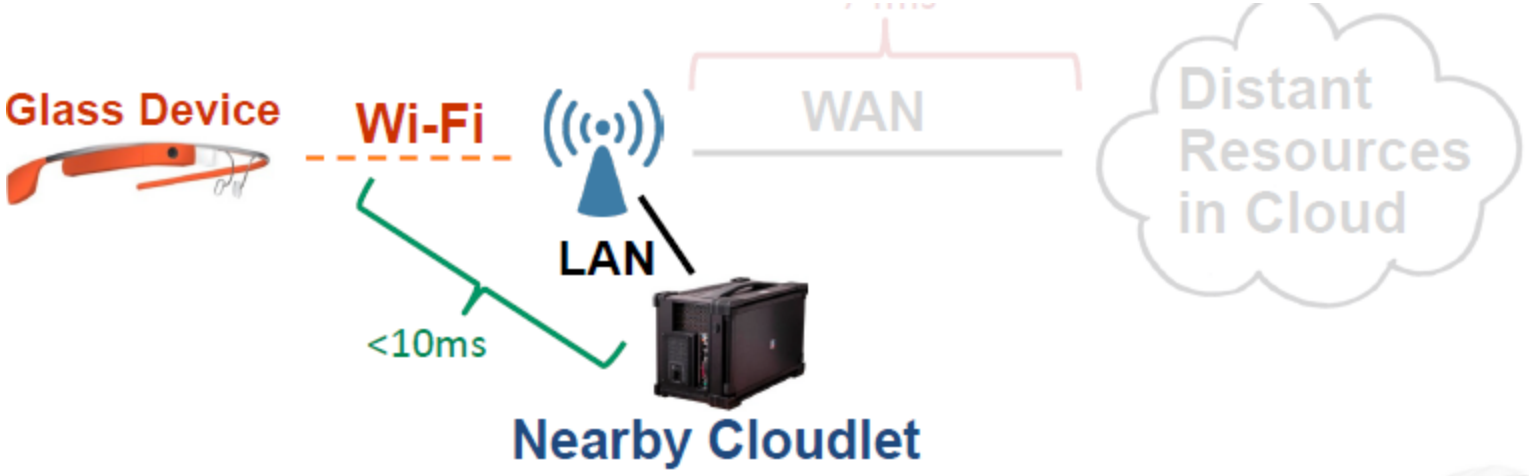
  optimal Amazon site ~74 ms

  heavy tailed distribution

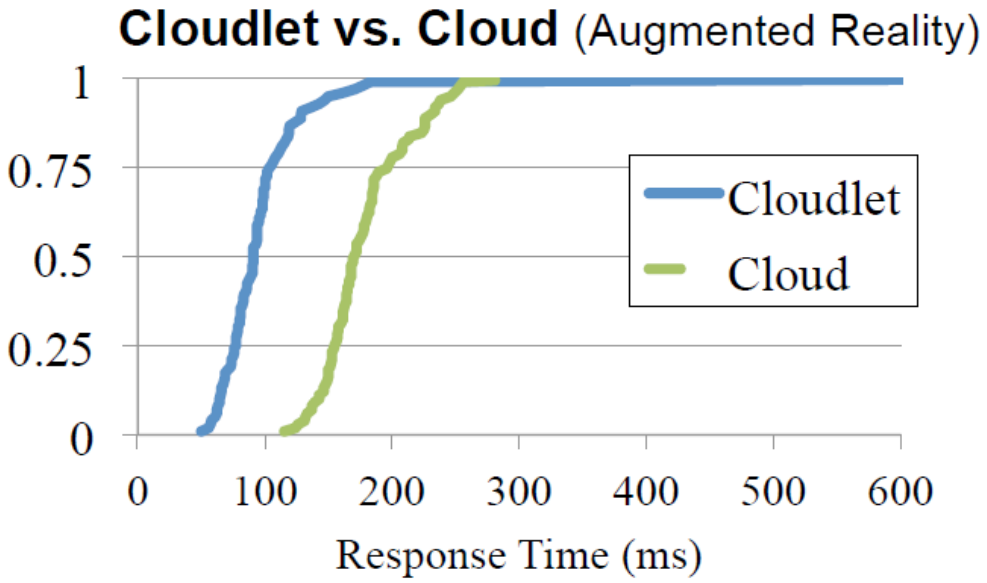
Solution 1: Crisp Interactive Response

- **Offload to cloudlet**

  data center in box
  bring cloud closer
  1-hop Wi-Fi access
  typical RTT < 10 ms



Exp. – Cloudlet Shortens Latency



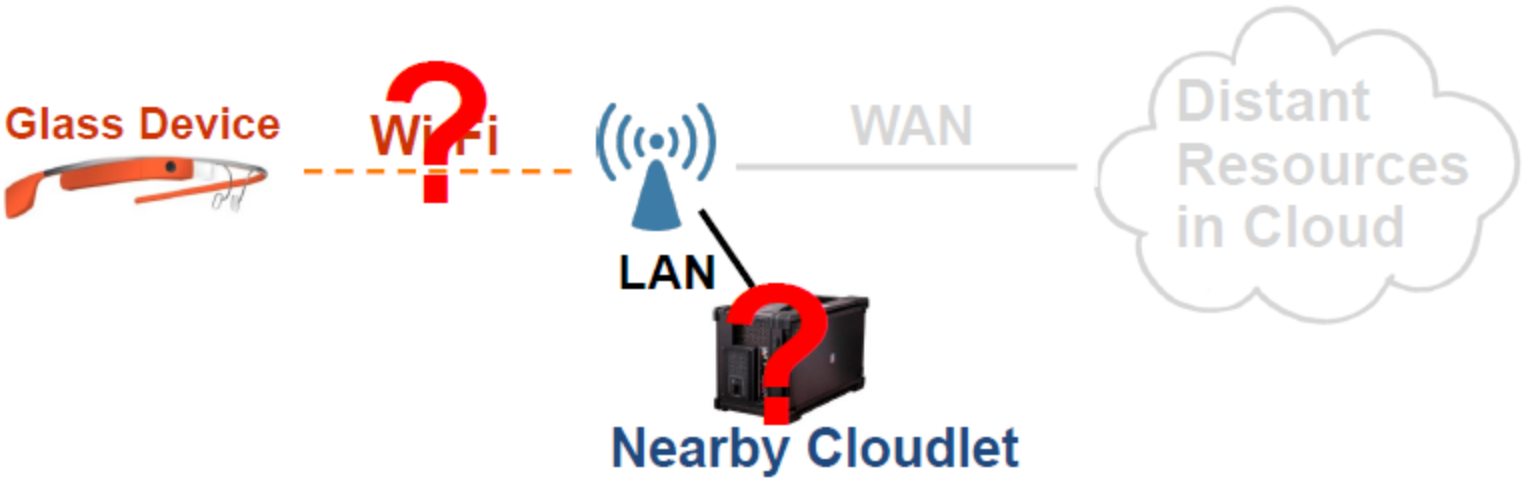
Cloudlet shortens response time

Challenge 2. Graceful Degradation of Services

What if offloading impossible?

Situation 1: No cloudlet

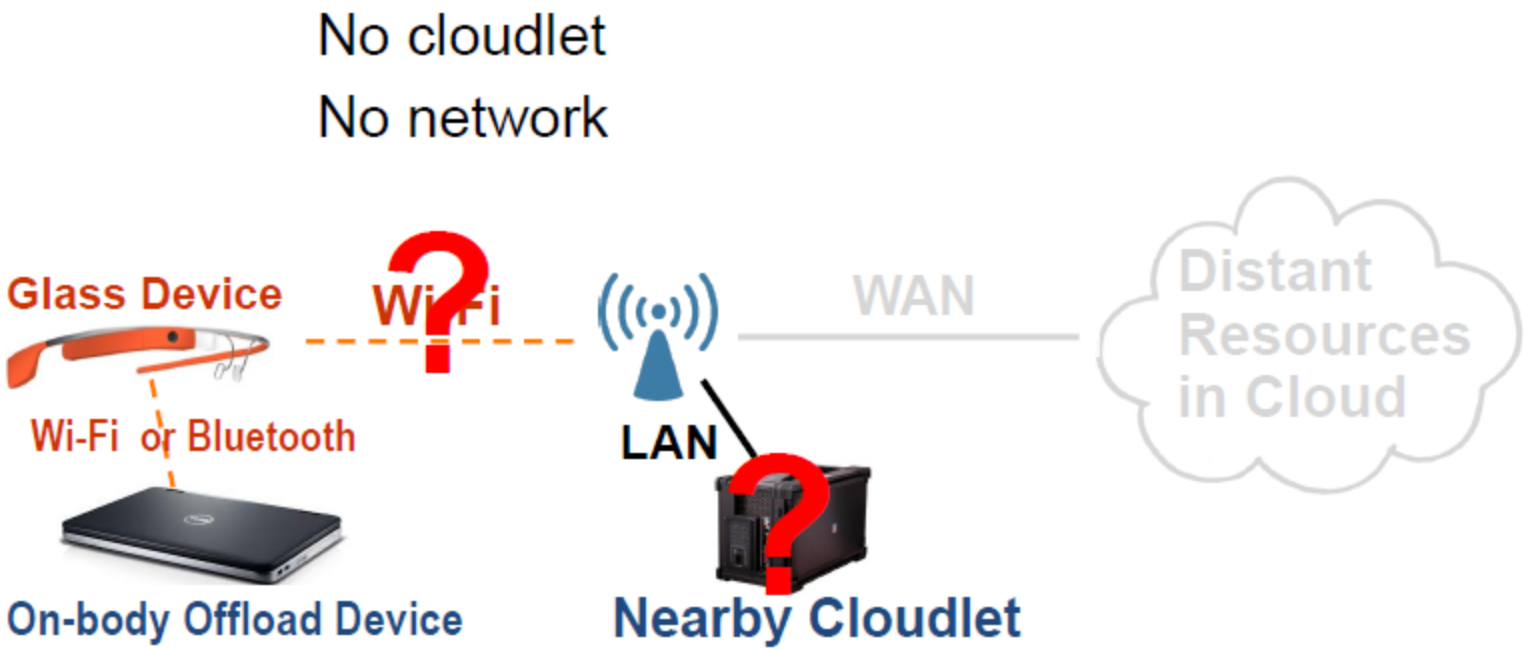
Situation 2: No network



Goal: still work during failures – with performance drop

Solution 2. Graceful Degradation of Services

Use fallback resources



Application-specific fidelity vs. Crispness & battery life

Challenge 2. Coarse-grain Parallelism

Face recognition
Object detection
Activity inference
OCR ……

Goal: reuse existing work, but…

- Programming languages are different

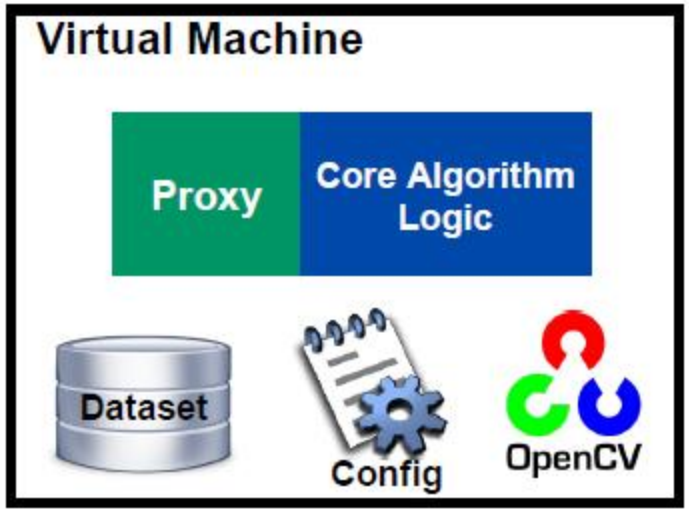
C Language Java python MATLAB

- Runtime systems are different (different OSes, closed-source, etc.)

Microsoft Windows Linux Mac OS

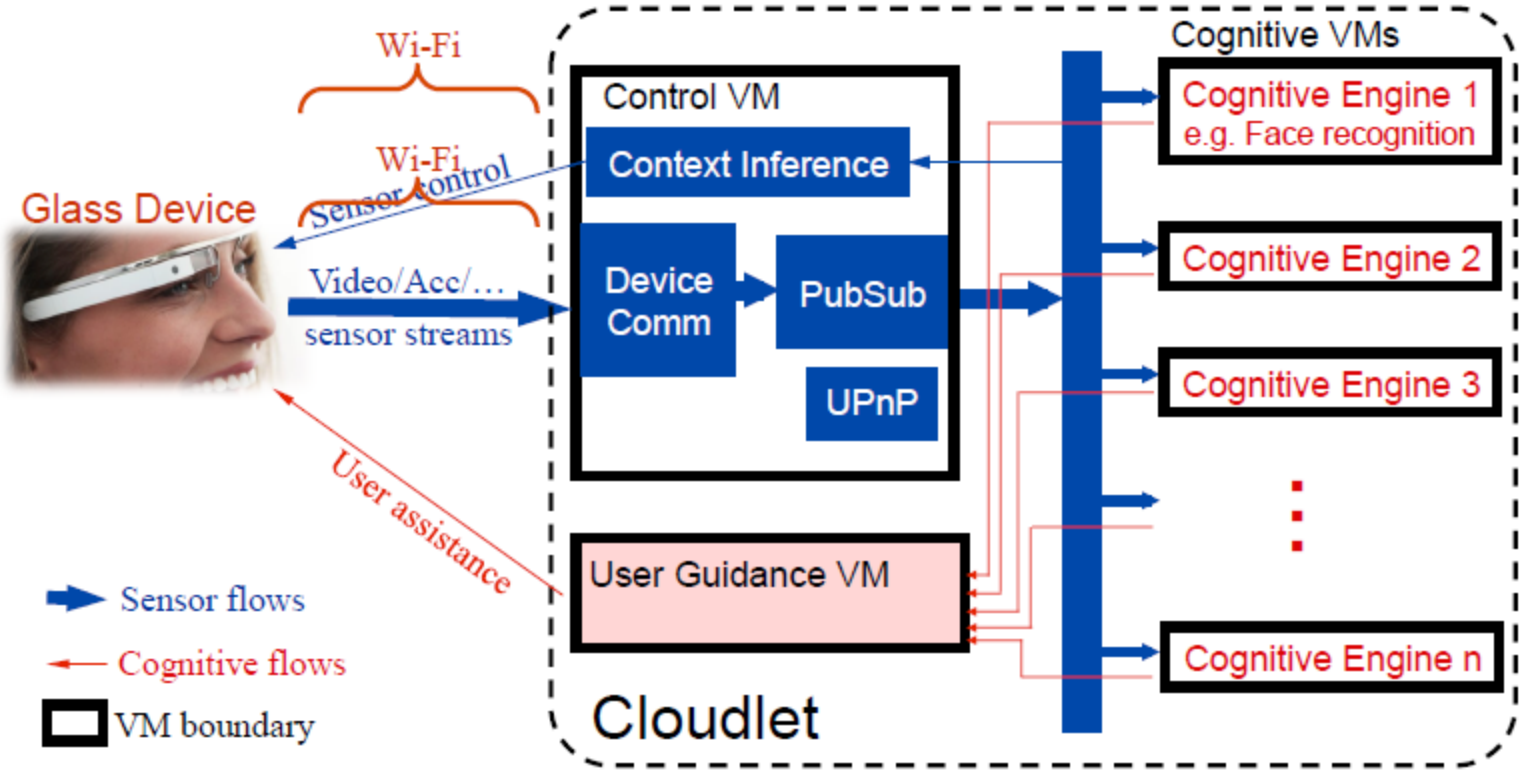
Solution 3. Coarse-grain Parallelism

VM Ensemble and PubSub Backbone



Solution 3. Coarse-grain Parallelism

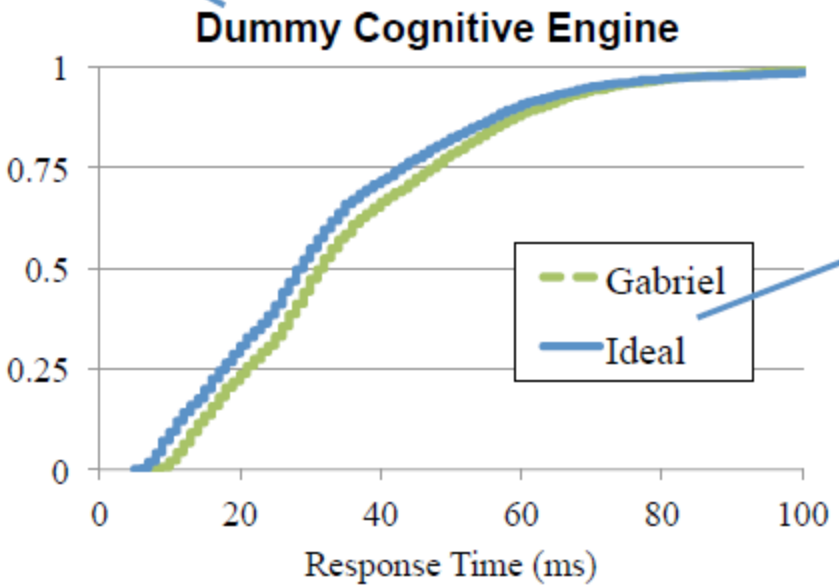
VM Ensemble and PubSub Backbone



Exp. – Gabriel Overhead

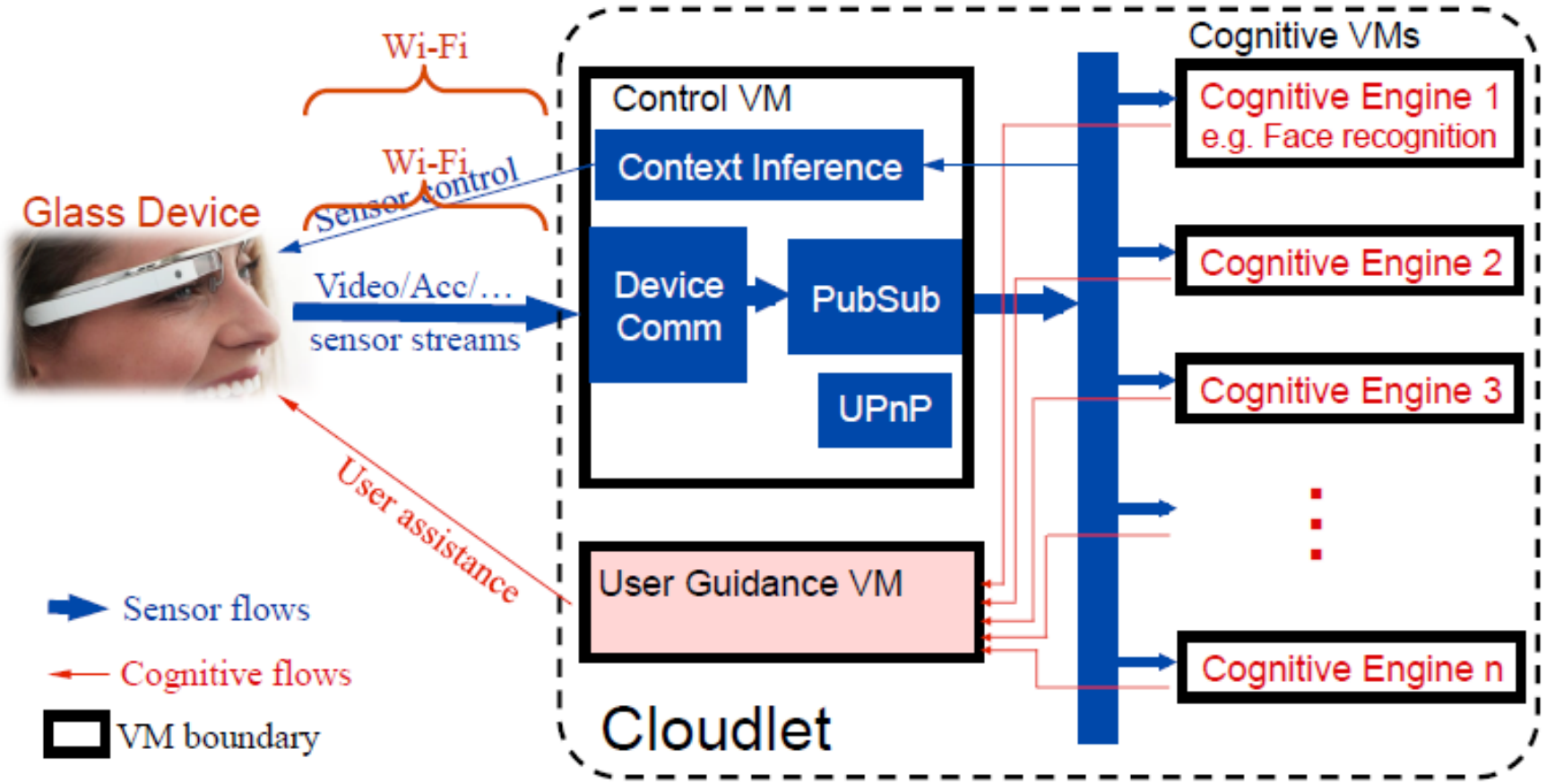
Gabriel Latency: ~50ms
Gabriel Overhead: ~3ms

Echos every image

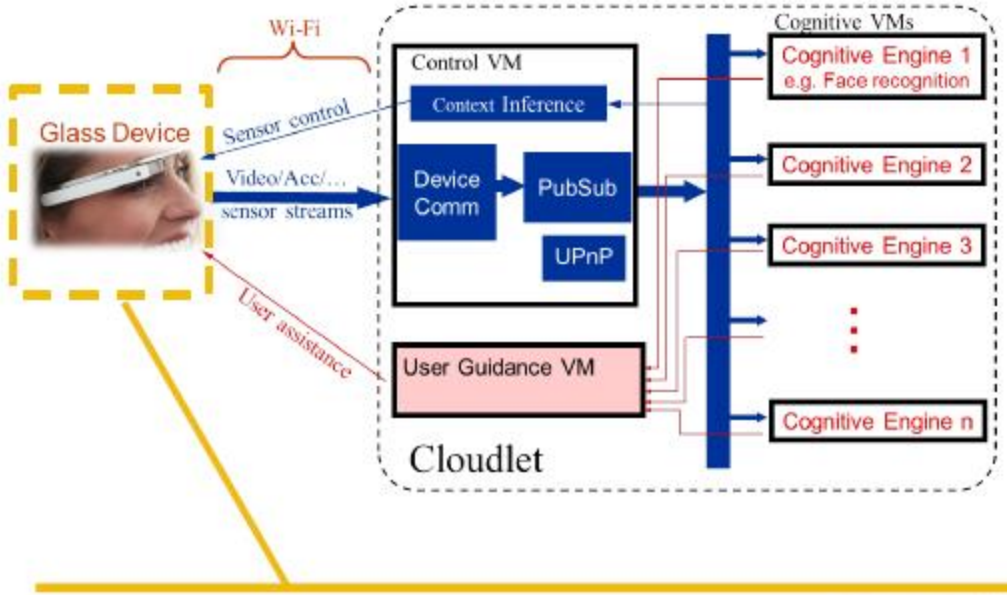


No VM, echos from host

System Architecture



Prototype Implementation



Prototype

Back-end Server

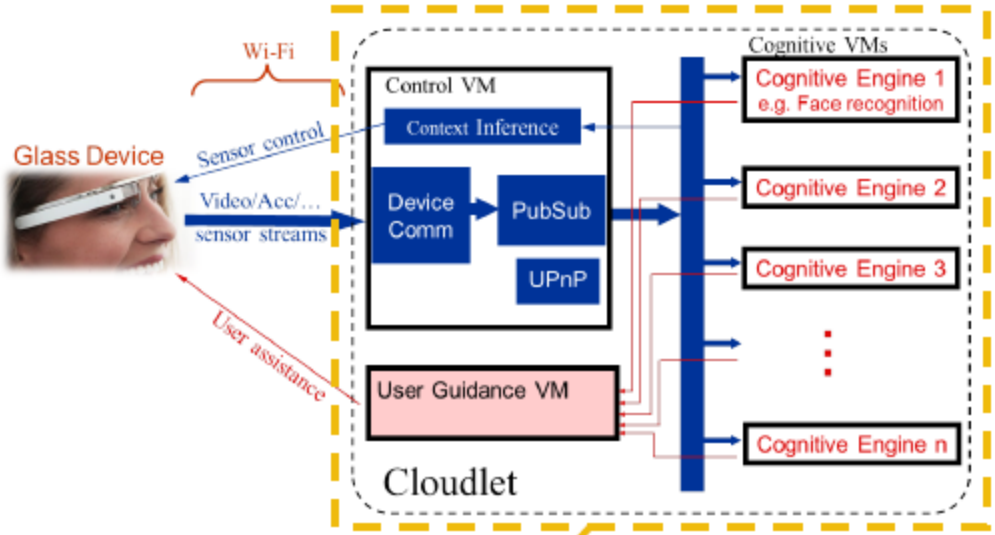
GDK Preview

TCP Connection

Speech Guidance

Ice pack to cool down Glass

Prototype Implementation



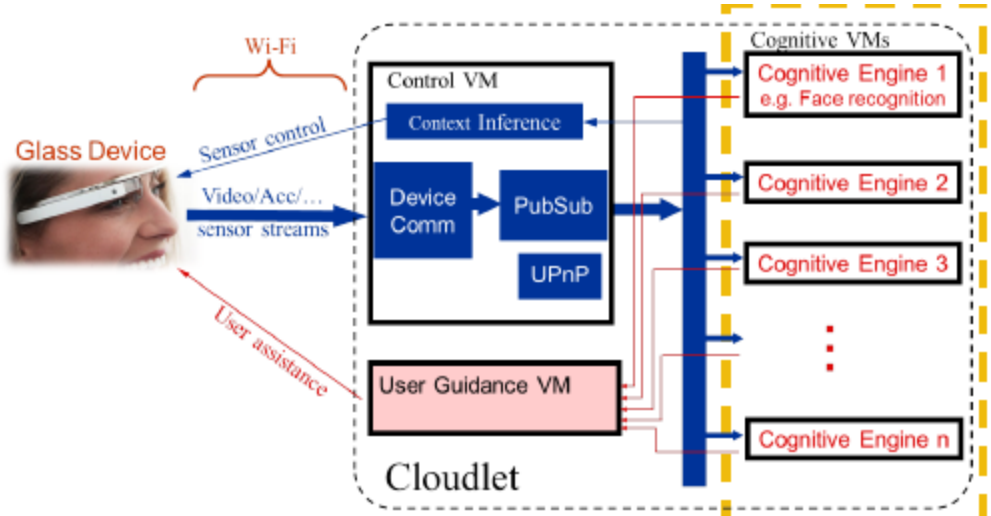
Prototype

Back-end Server

Cloudlet: 4 advanced desktop machines

Running OpenStack – Virtualized Cloud Computing Platform

Prototype Implementation



Prototype

Cognitive Engines

Windows Linux

C Language C++ python

- Face Recognition
- Object Recognition (1. MOPED 2. STF)
- OCR (1. Tesseract 2. VeryPDF) — Commercial Product
- Motion Classifier
- Augmented Reality
- Activity Detection — Based on Accelerometer

Exp. – Full System Performance

Cognitive Engines are slower

Cognitive Engine	FPS	Response time (ms)					Glass Life
		1%	10%	50%	90%	99%	
Face Recognition	4.4	196	389	659	929	1175	
Object (MOPED)	1.6	877	962	1207	1647	2118	
Object (STF)	0.4	4202	4371	4609	5055	5684	
OCR (Open)	14.4	29	41	87	147	511	~1 hour
OCR (Comm)	2.3	394	435	522	653	1021	
Motion Classifier	14.0	126	152	199	260	649	
Augmented Reality	14.1	48	72	126	192	498	

Exp. – Full System Performance

Cognitive Engines require different FPS

Cognitive Engine	FPS	Response time (ms)					Glass Life
		1%	10%	50%	90%	99%	
Face Recognition	4.4	196	389	659	929	1175	
Object (MOPED)	1.6	877	962	1207	1647	2118	
Object (STF)	0.4	4202	4371	4609	5055	5684	
OCR (Open)	14.4	29	41	87	147	511	~1 hour
OCR (Comm)	2.3	394	435	522	653	1021	
Motion Classifier	14.0	126	152	199	260	649	
Augmented Reality	14.1	48	72	126	192	498	

More in Paper

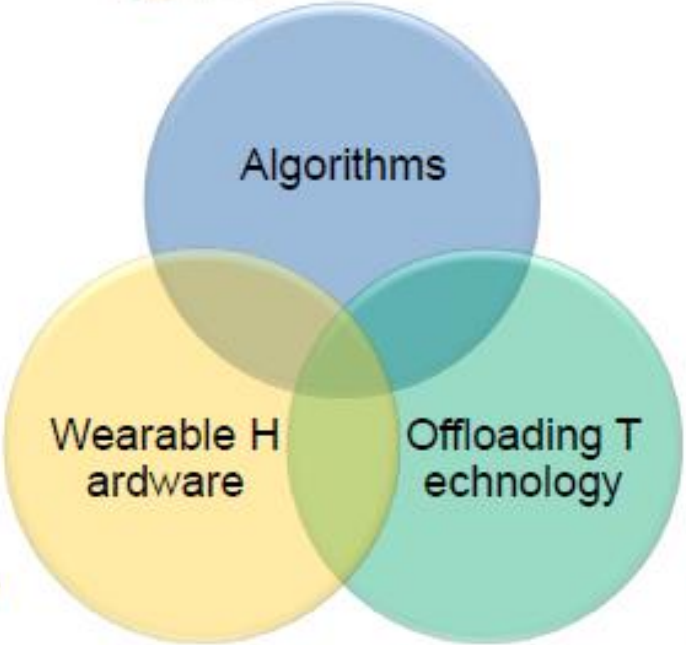
1. Token-based flow control improves response time a lot

2. Gabriel supports multi-VM parallelism

3. Tradeoff between fidelity reduction and crisp user interaction

Conclusion & Future Work

Speed improvement needed



Longer battery, better thermal dissipation

Cloudlets are helpful, need good biz. model