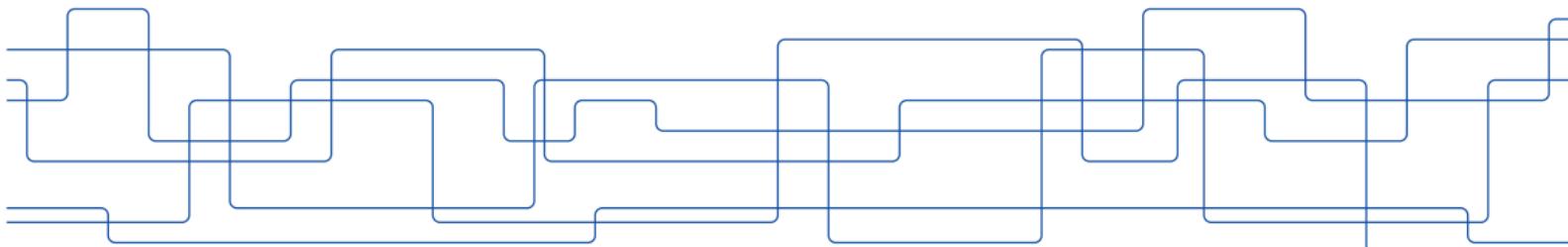




An Experimental Approach to Benchmarking Human-in-the-Loop Applications

Manuel Olguín Muñoz

ISE Internal Seminar, Thursday March 28 2019



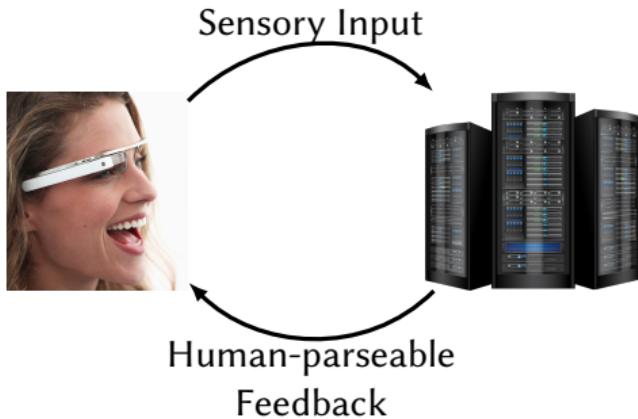
Acknowledgements

Part of an ongoing collaboration with the Elijah Group at Carnegie Mellon University, led by Prof. Mahadev Satyanarayanan.



This work was recently presented at HotMobile'19 [1].





Studying Human-in-the-Loop Applications

Need to understand and optimize these applications:

- ▶ How do they interact with each other?
- ▶ How do they interact with infrastructure?
- ▶ How do they scale?

With which methodology can we study these behaviors?

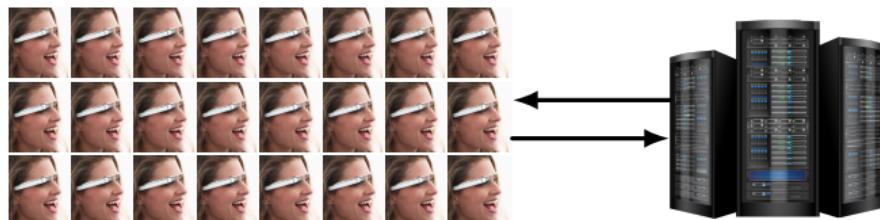


Studying Human-in-the-Loop Applications

Need to understand and optimize these applications:

- ▶ How do they interact with each other?
- ▶ How do they interact with infrastructure?
- ▶ How do they scale?

With which methodology can we study these behaviors?

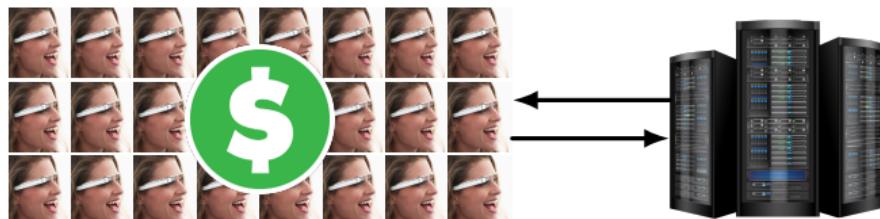


Studying Human-in-the-Loop Applications

Need to understand and optimize these applications:

- ▶ How do they interact with each other?
- ▶ How do they interact with infrastructure?
- ▶ How do they scale?

With which methodology can we study these behaviors?



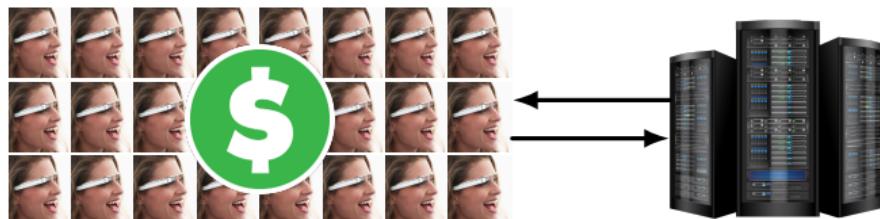
Costly

Studying Human-in-the-Loop Applications

Need to understand and optimize these applications:

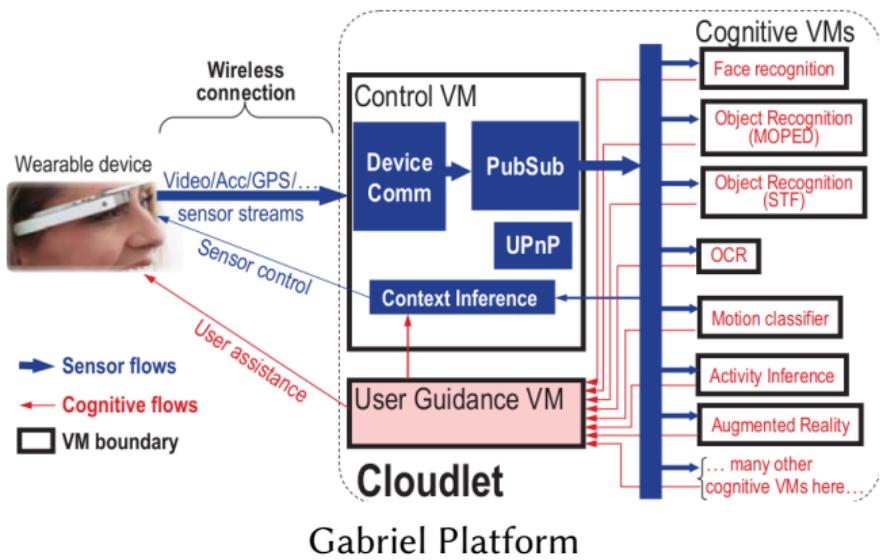
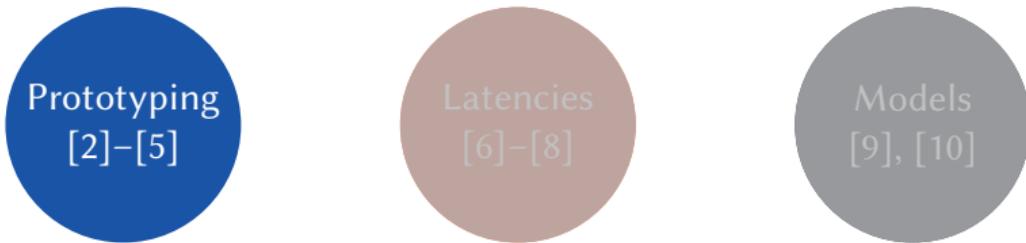
- ▶ How do they interact with each other?
- ▶ How do they interact with infrastructure?
- ▶ How do they scale?

With which methodology can we study these behaviors?

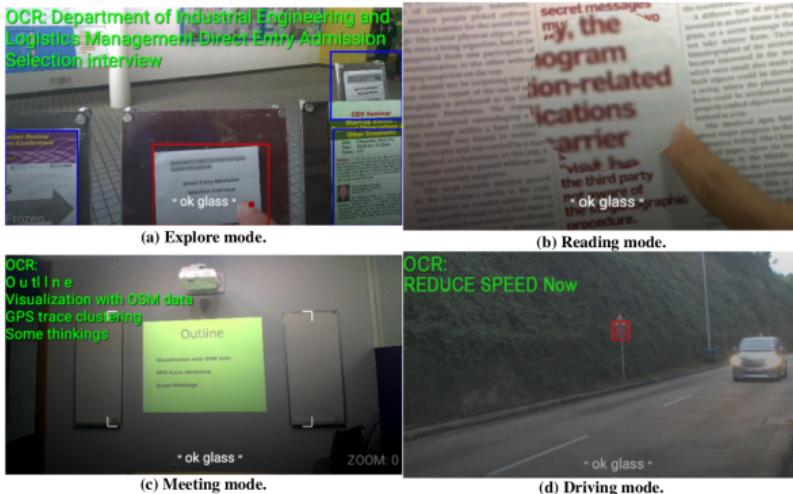
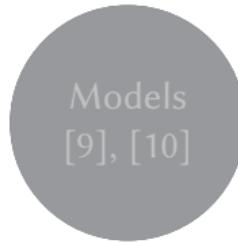
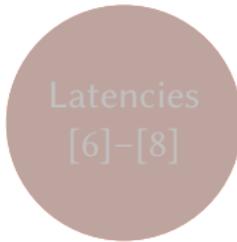
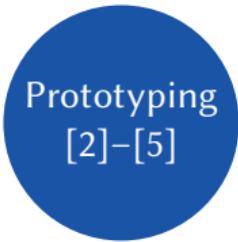


Costly, poor repeatability

Previous & Related Work



Previous & Related Work



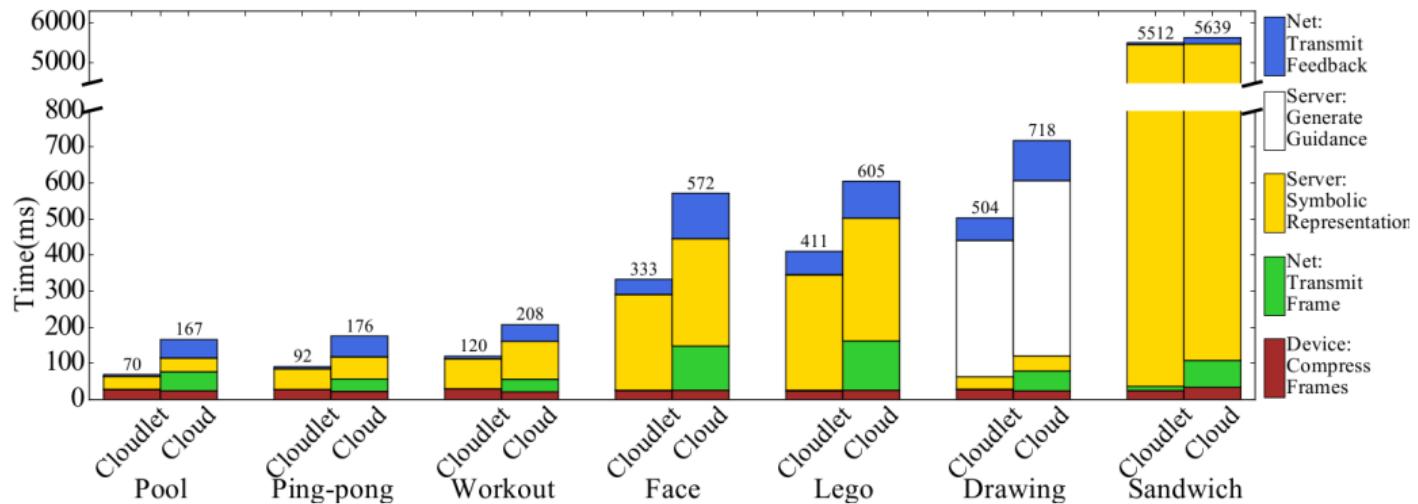
Hyperion Assistant

Previous & Related Work

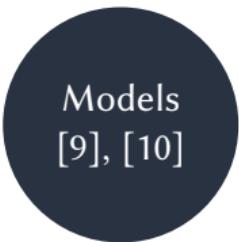
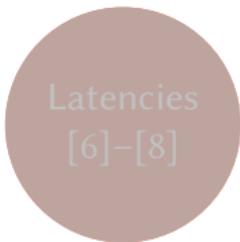
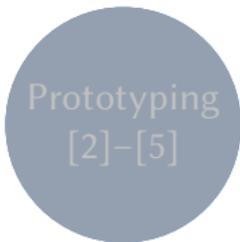
Prototyping
[2]–[5]

Latencies
[6]–[8]

Models
[9], [10]



Previous & Related Work



- ▶ H. Al-Zubaidy *et al.*, “Performance of in-network processing for visual analysis in wireless sensor networks,” in *Proceedings of the IFIP Networking Conference*, ser. IFIP NETWORKING’15, 2015
- ▶ S. Schiessl *et al.*, “Finite-length coding in edge computing scenarios,” in *Proceedings of the International Workshop on Smart Antennas*, ser. ITG WSA ’17, 2017

Our Contributions

- ▶ A methodology for benchmarking human-in-the-loop applications.

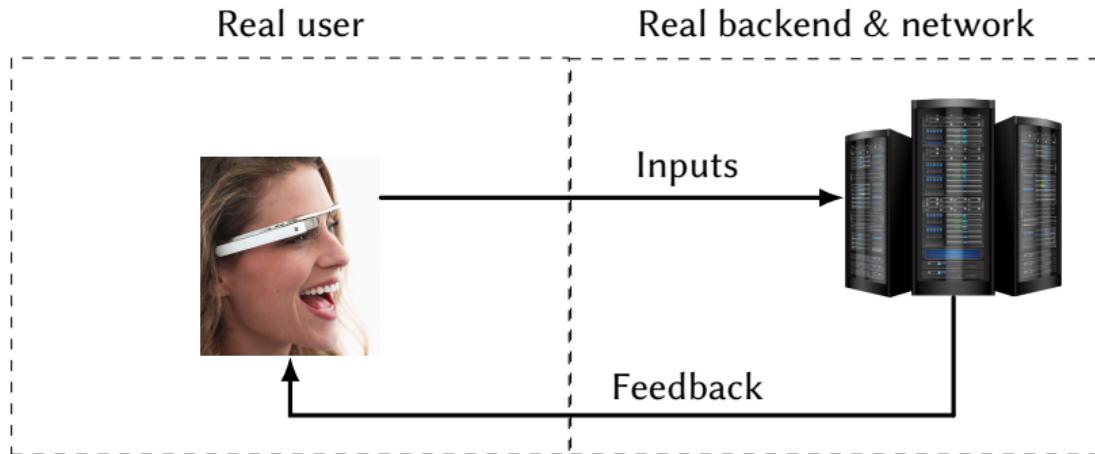
Our Contributions

- ▶ A methodology for benchmarking human-in-the-loop applications.
- ▶ EdgeDroid: A benchmarking tool-suite.

Our Contributions

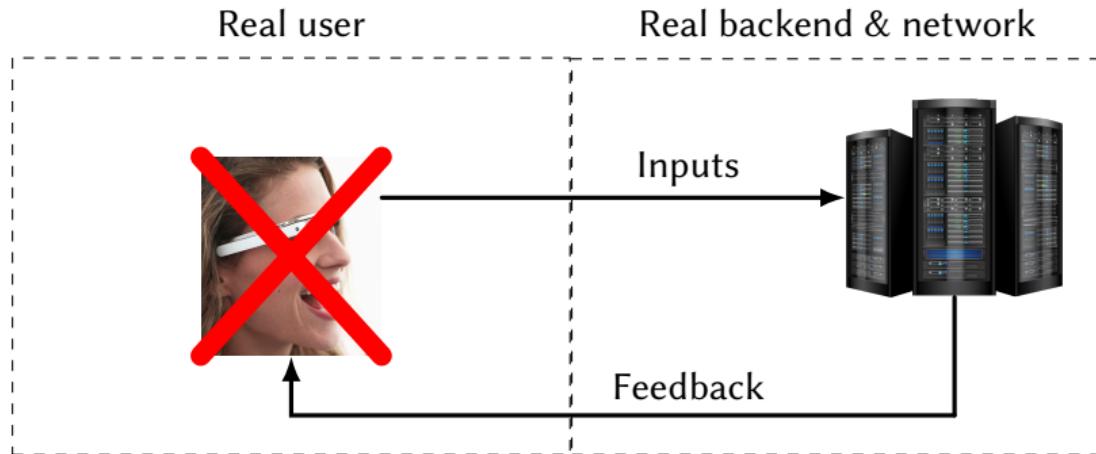
- ▶ A methodology for benchmarking human-in-the-loop applications.
- ▶ EdgeDroid: A benchmarking tool-suite.
- ▶ Experiments and measurements which show the effectiveness of the approach.

Approach



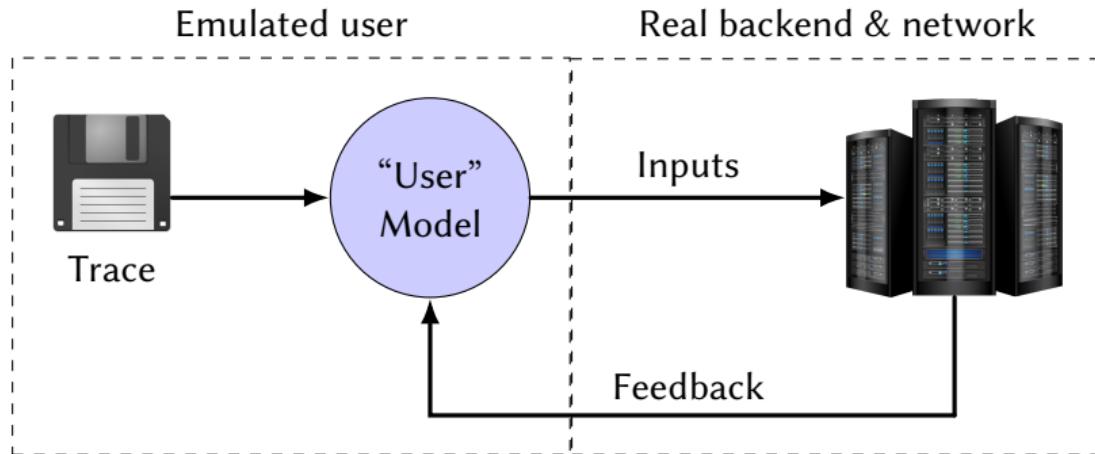
Benchmarking human-in-the-loop applications is HARD

Approach



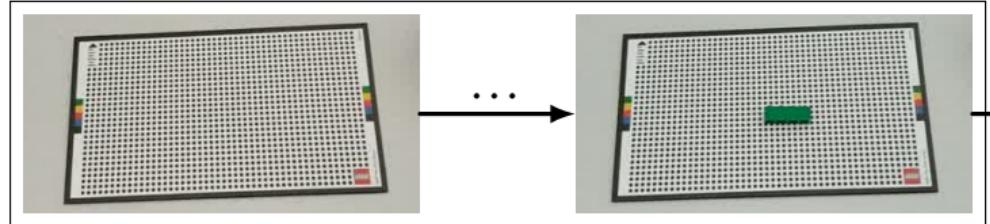
What if we could do away with the human users?

Approach



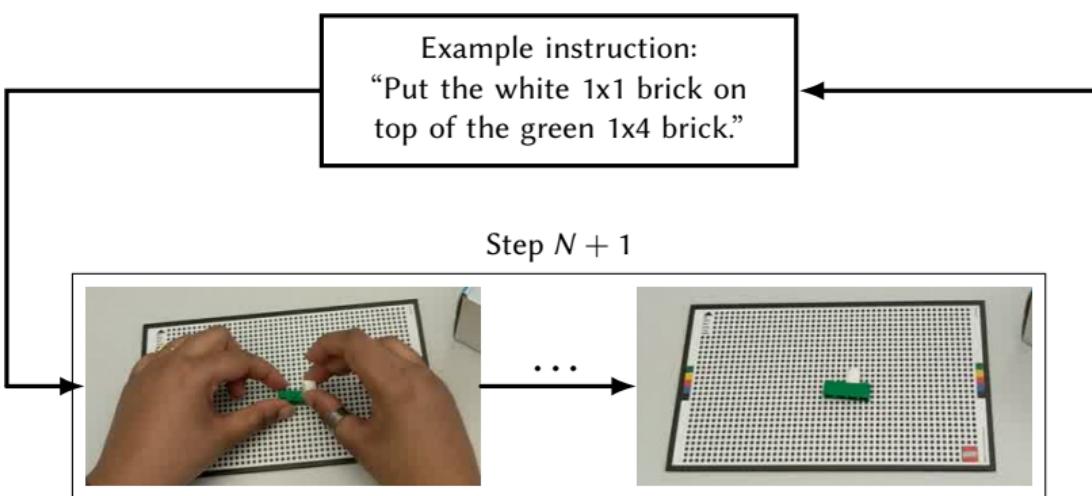
What if we could do away with the human users?
Repeatable, scalable!

Example: Task Guidance Wearable Cognitive Assitance, LEGO [2]

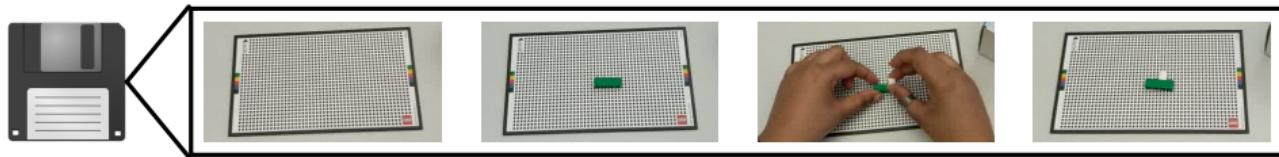
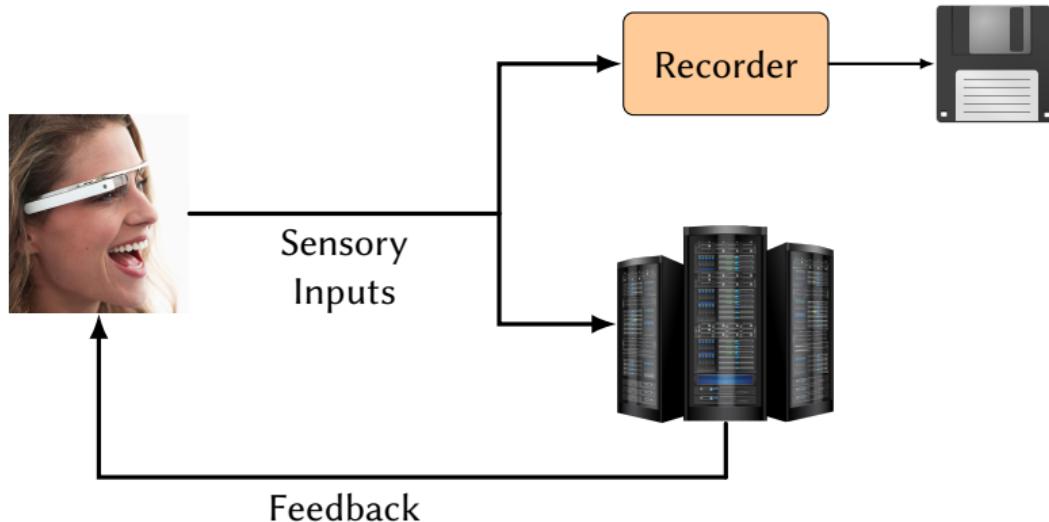


Example instruction:
“Put the white 1x1 brick on
top of the green 1x4 brick.”

Step $N + 1$



Tracing



Trace Replay

Non-trivial Challenge

- ▶ Changes in system responsiveness require adapting trace.
- ▶ System delays affect user behavior as well.

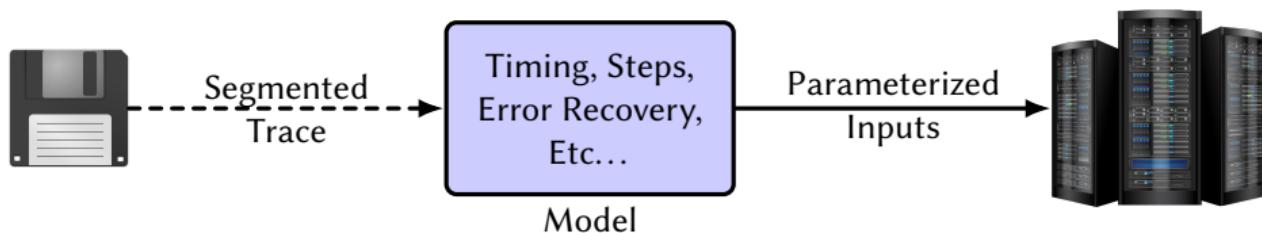
Trace Replay

Non-trivial Challenge

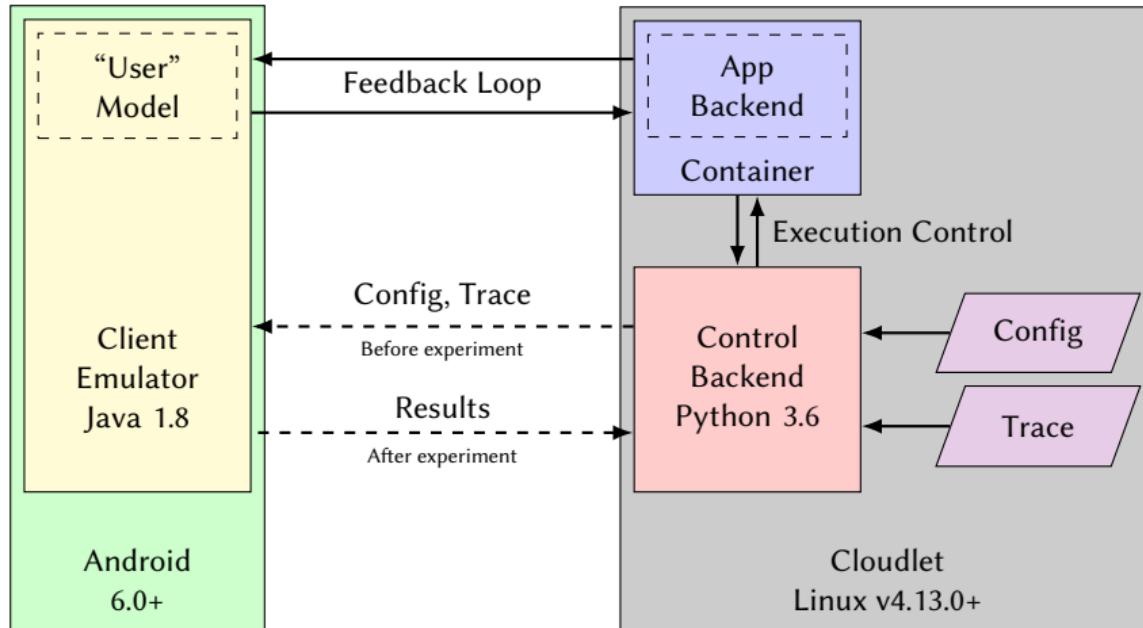
- ▶ Changes in system responsiveness require adapting trace.
- ▶ System delays affect user behavior as well.

Our Approach

- ▶ Segment trace into logical “steps”.
- ▶ Controlled replay of steps.

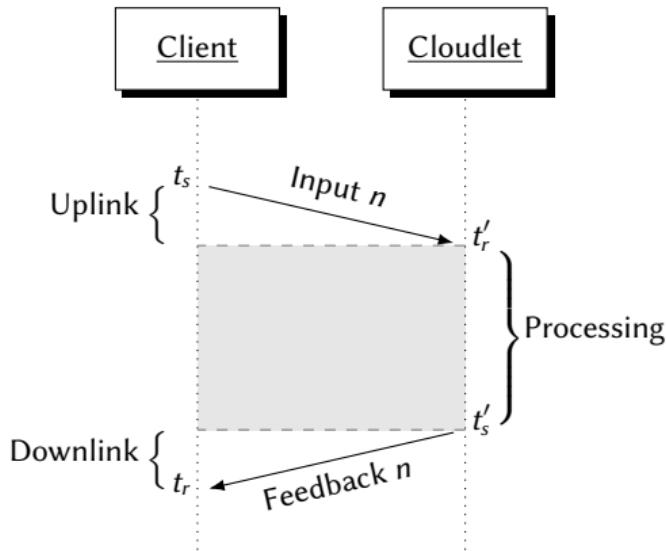


Implementation: EdgeDroid



<https://github.com/molguin92/EdgeDroid>

Timestamping



Clocks are synchronized previous to the experiment.

Timestamps at key points to obtain:

$$\Delta T_{\text{up}} = t_r' - t_s \quad (1)$$

$$\Delta T_{\text{proc}} = t_s' - t_r' \quad (2)$$

$$\Delta T_{\text{down}} = t_r - t_s' \quad (3)$$

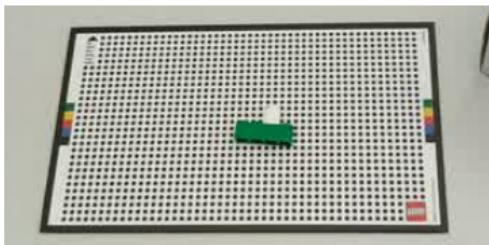
$$\Delta T_{\text{total}} = \Delta T_{\text{up}} + \Delta T_{\text{proc}} + \Delta T_{\text{down}} = t_r - t_s \quad (4)$$

Evaluation

Key purpose:
Demonstrate utility of EdgeDroid.

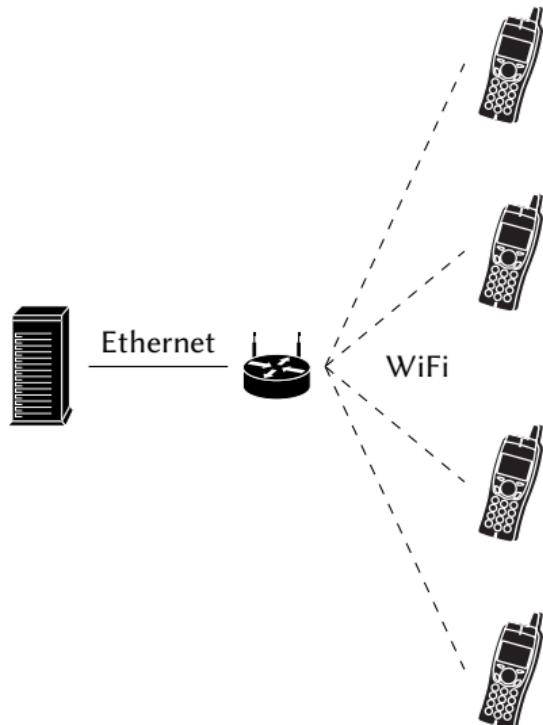
Evaluation: Setup

Application & Scenarios



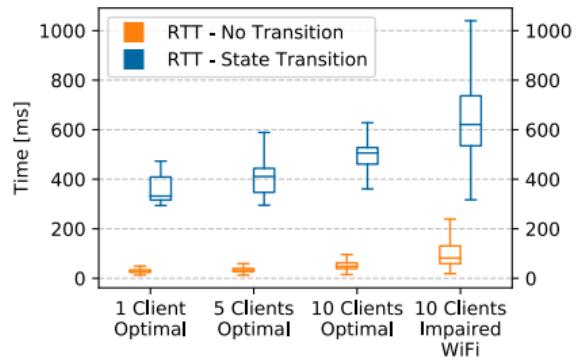
LEGO Assistant

- ▶ Three *optimal* scenarios with 1, 5 and 10 devices.
- ▶ Weakened wireless link with 10 devices.
- ▶ KPI: Round-Trip Time (RTT).

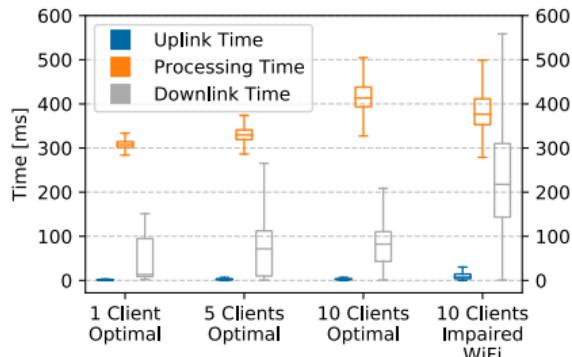


Use Cases

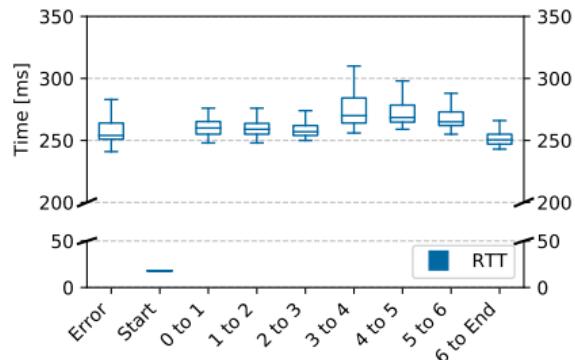
State change vs. no state change.



Times by pipeline segments.



RTT by task step.



Future Work

- ▶ Extending EdgeDroid
 - ▶ Characterizing human behavior
 - ▶ Extending to other types of applications
- ▶ Characterizing Control Performance on Edge Computing Infrastructure

Characterizing human behavior...

...in the presence of delays in human-in-the-loop applications.

Open research question: how do delays affect users?

Task performance

of Errors

Task completion time



Confusion

Annoyance

Fatigue

Conclusions

Future Work

- ▶ User Model.
- ▶ Other types of Applications.

Summary

- ▶ Need to study the scaling of Human-in-the-Loop applications.
 - ▶ Difficult due to human users.
- ▶ Methodology + tool suite for benchmarking:
 - ▶ **EdgeDroid**
 - ▶ Trace based.
 - ▶ Model of human behavior.
- ▶ Results which show the utility of EdgeDroid.



Thank you.

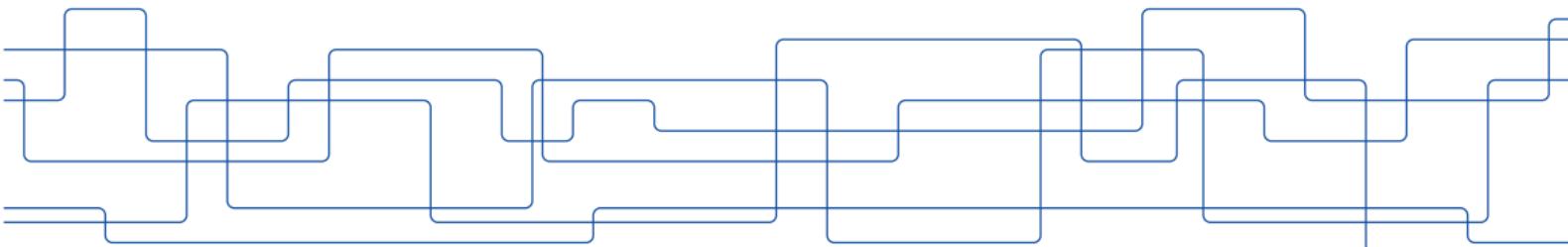
Contact

Manuel Olguín Muñoz

Division of Information Science and Engineering
KTH EECS
Malvinas väg 10, 100-44 Stockholm, SWEDEN

Email: molguin@kth.se

Website: <https://olguin.se>



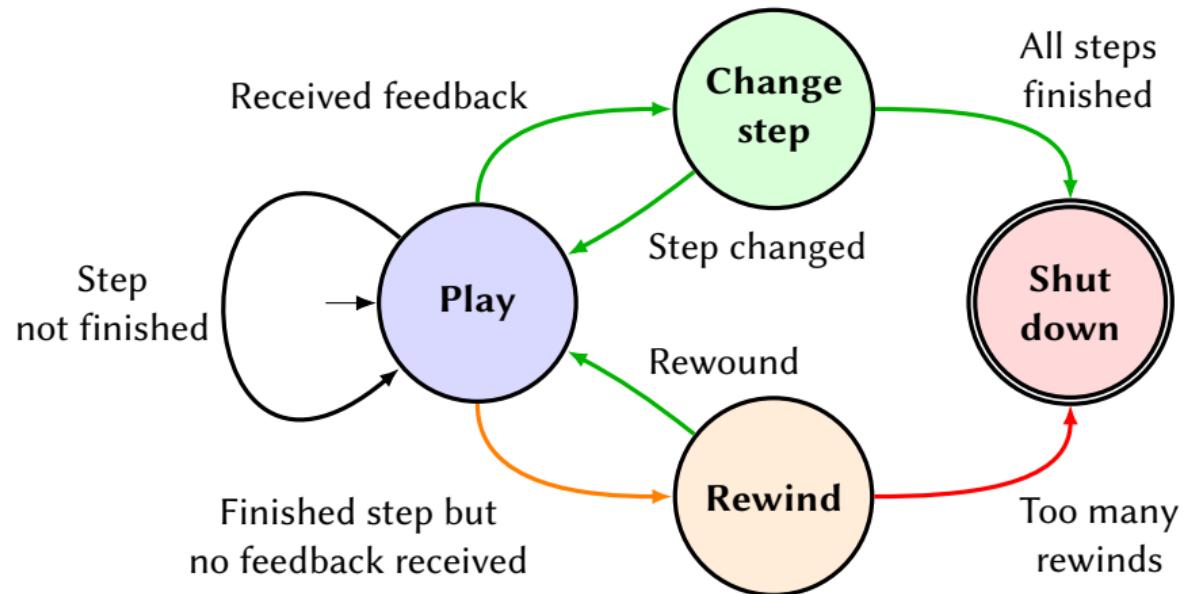
Requirements

- ▶ Generate realistic, high-dimensional, real-time inputs.
- ▶ Correctly and realistically react to feedback.
- ▶ KPI: Delays.



**Trace of pre-recorded inputs
& a model of user behavior**

User Model



Future work: more elaborate models.

References I

- [1] M. O. J. Olguín Muñoz *et al.*, “EdgeDroid: An Experimental Approach to Benchmarking Human-in-the-Loop Applications,” in *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications*, ser. HotMobile ’19, Santa Cruz, CA, USA: ACM, 2019, pp. 93–98, ISBN: 978-1-4503-6273-3. doi: 10.1145/3301293.3302353. [Online]. Available: <http://doi.acm.org/10.1145/3301293.3302353>.
- [2] K. Ha *et al.*, “Towards wearable cognitive assistance,” in *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys ’14, Bretton Woods, New Hampshire, USA: ACM, 2014, pp. 68–81, ISBN: 978-1-4503-2793-0. doi: 10.1145/2594368.2594383. [Online]. Available: <http://doi.acm.org/10.1145/2594368.2594383>.
- [3] Z. Chen *et al.*, “Early implementation experience with wearable cognitive assistance applications,” in *Proceedings of the 2015 Workshop on Wearable Systems and Applications*, ser. WearSys ’15, Florence, Italy: ACM, 2015, pp. 33–38, ISBN: 978-1-4503-3500-3. doi: 10.1145/2753509.2753517. [Online]. Available: <http://doi.acm.org/10.1145/2753509.2753517>.

References II

- [4] D. Chatzopoulos *et al.*, “Hyperion: A wearable augmented reality system for text extraction and manipulation in the air,” in *Proceedings of the 8th ACM on Multimedia Systems Conference*, ser. MMSys’17, Taipei, Taiwan: ACM, 2017, pp. 284–295, ISBN: 978-1-4503-5002-0. doi: 10.1145/3083187.3084017. [Online]. Available: <http://doi.acm.org/10.1145/3083187.3084017>.
- [5] S. Jalaliniya *et al.*, “Designing wearable personal assistants for surgeons: An egocentric approach,” *IEEE Pervasive Computing*, vol. 14, no. 3, pp. 22–31, 2015, ISSN: 1536-1268. doi: 10.1109/MPRV.2015.61.
- [6] Z. Chen *et al.*, “An empirical study of latency in an emerging class of edge computing applications for wearable cognitive assistance,” in *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*, ser. SEC ’17, San Jose, California: ACM, 2017, 14:1–14:14, ISBN: 978-1-4503-5087-7. doi: 10.1145/3132211.3134458. [Online]. Available: <http://doi.acm.org/10.1145/3132211.3134458>.
- [7] J. Dolezal *et al.*, “Performance evaluation of computation offloading from mobile device to the edge of mobile network,” in *2016 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2016, pp. 1–7. doi: 10.1109/CSCN.2016.7785153.
- [8] D. Chatzopoulos *et al.*, “Mobile augmented reality survey: From where we are to where we go,” *IEEE Access*, vol. 5, pp. 6917–6950, 2017, ISSN: 2169-3536. doi: 10.1109/ACCESS.2017.2698164.

References III

- [9] H. Al-Zubaidy *et al.*, “Performance of in-network processing for visual analysis in wireless sensor networks,” in *Proceedings of the IFIP Networking Conference*, ser. IFIP NETWORKING’15, 2015.
- [10] S. Schiessl *et al.*, “Finite-length coding in edge computing scenarios,” in *Proceedings of the International Workshop on Smart Antennas*, ser. ITG WSA ’17, 2017.
- [11] M. Satyanarayanan *et al.*, “The case for VM-based cloudlets in mobile computing,” *IEEE Pervasive Computing*, vol. 8, no. 4, 2009.
- [12] J. Flinn, “Cyber foraging: Bridging mobile and cloud computing,” *Synthesis Lectures on Mobile and Pervasive Computing*, vol. 7, no. 2, pp. 1–103, 2012.
- [13] K. Sasaki *et al.*, “Vehicle control system coordinated between cloud and mobile edge computing,” in *2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, 2016, pp. 1122–1127. doi: [10.1109/SICE.2016.7749210](https://doi.org/10.1109/SICE.2016.7749210).
- [14] ——, “Layered vehicle control system coordinated between multiple edge servers,” in *2017 IEEE Conference on Network Softwarization (NetSoft)*, 2017, pp. 1–5. doi: [10.1109/NETSOFT.2017.8004199](https://doi.org/10.1109/NETSOFT.2017.8004199).
- [15] T. Bittmann, “The edge will eat the cloud,” *Gartner Research*, no. G00338633, 2017.
- [16] K. Kumar *et al.*, “Cloud computing for mobile users: Can offloading computation save energy?” *IEEE Computer*, vol. 43, no. 4, pp. 51–56, 2010.

References IV

- [17] E. Cuervo *et al.*, “Maui: Making smartphones last longer with code offload,” in *Proceedings of the International Conference on Mobile Systems, Applications, and Services*, ser. ACM MOBISYS’10, 2010.
- [18] K. Ha *et al.*, “The impact of mobile multimedia applications on data center consolidation,” in *2013 IEEE International Conference on Cloud Engineering (IC2E)*, 2013, pp. 166–176. doi: 10.1109/IC2E.2013.17.
- [19] K. Ha *et al.*, “Just-in-time provisioning for cyber foraging,” in *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys ’13, Taipei, Taiwan: ACM, 2013, pp. 153–166, ISBN: 978-1-4503-1672-9. doi: 10.1145/2462456.2464451. [Online]. Available: <http://doi.acm.org/10.1145/2462456.2464451>.
- [20] (2018). Docker, [Online; accessed 14. Aug. 2018], [Online]. Available: <https://www.docker.com>.
- [21] (2018). Network Time Protocol, [Online; accessed 24. Sep. 2018], [Online]. Available: <https://www.eecis.udel.edu/~mills/ntp/html/index.html>.
- [22] (2018). TOML, [Online; accessed 25. Sep. 2018], [Online]. Available: <https://github.com/toml-lang/toml>.

References V

- [23] K. Kim *et al.*, “Workload synthesis: Generating benchmark workloads from statistical execution profile,” in *2014 IEEE International Symposium on Workload Characterization (IISWC)*, 2014, pp. 120–129. doi: 10.1109/IISWC.2014.6983051.
- [24] E. Deniz *et al.*, “Minime: Pattern-aware multicore benchmark synthesizer,” *IEEE Transactions on Computers*, vol. 64, no. 8, pp. 2239–2252, 2015, ISSN: 0018-9340. doi: 10.1109/TC.2014.2349522.
- [25] M. Olguín *et al.*, “Demo: Scaling on the Edge – A Benchmarking Suite for Human-in-the-Loop Applications,” in *Proceedings of The Third ACM/IEEE Symposium on Edge Computing*, ser. SEC ’18, Accepted Submission, Extended Abstract, 2018. [Online]. Available: <https://olguin.se/files/demo-scaling-edge.pdf>.