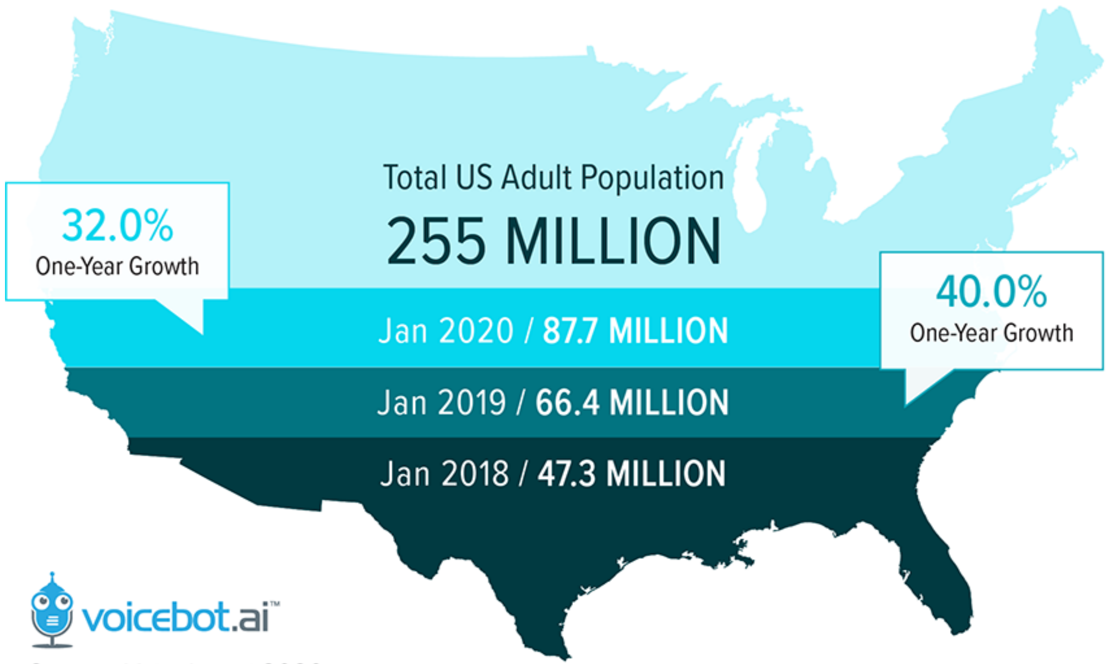
# *CHA*: A ***C***aching Framework for ***H***ome-based Voice ***A***ssistant Systems

Lanyu Xu[1], Arun Iyengar[2], Weisong Shi[1]

[1]Wayne State University
[2]IBM T.J. Watson Research Center

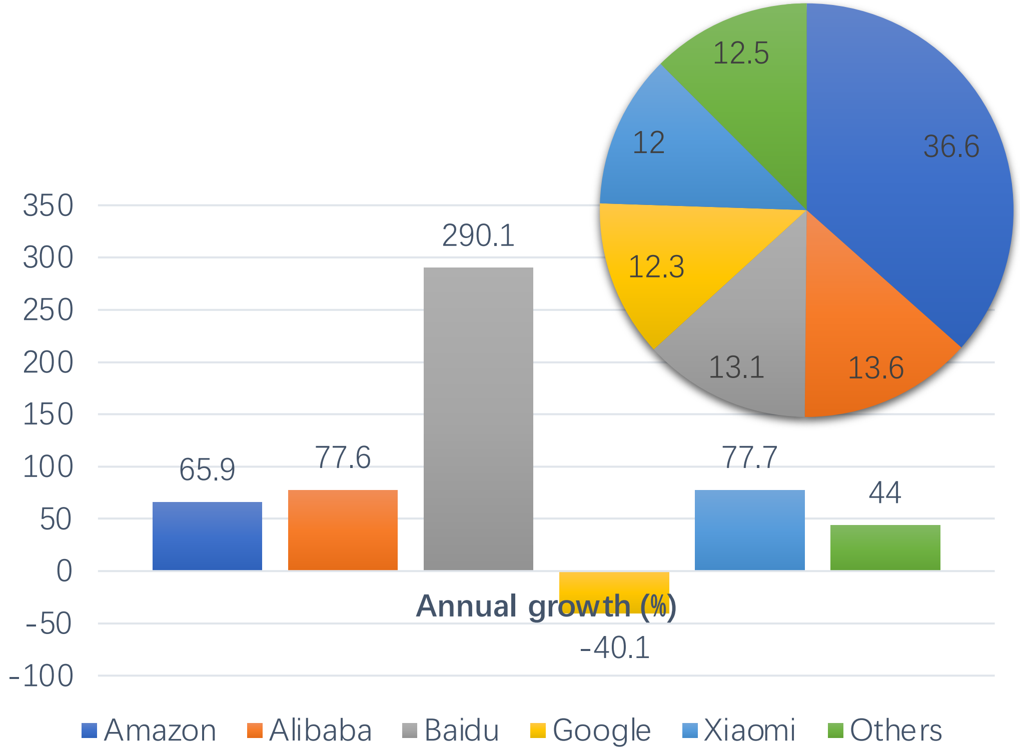
# Introduction: Smart Speaker

U.S. Adult Smart Speaker Installed Base January 2020



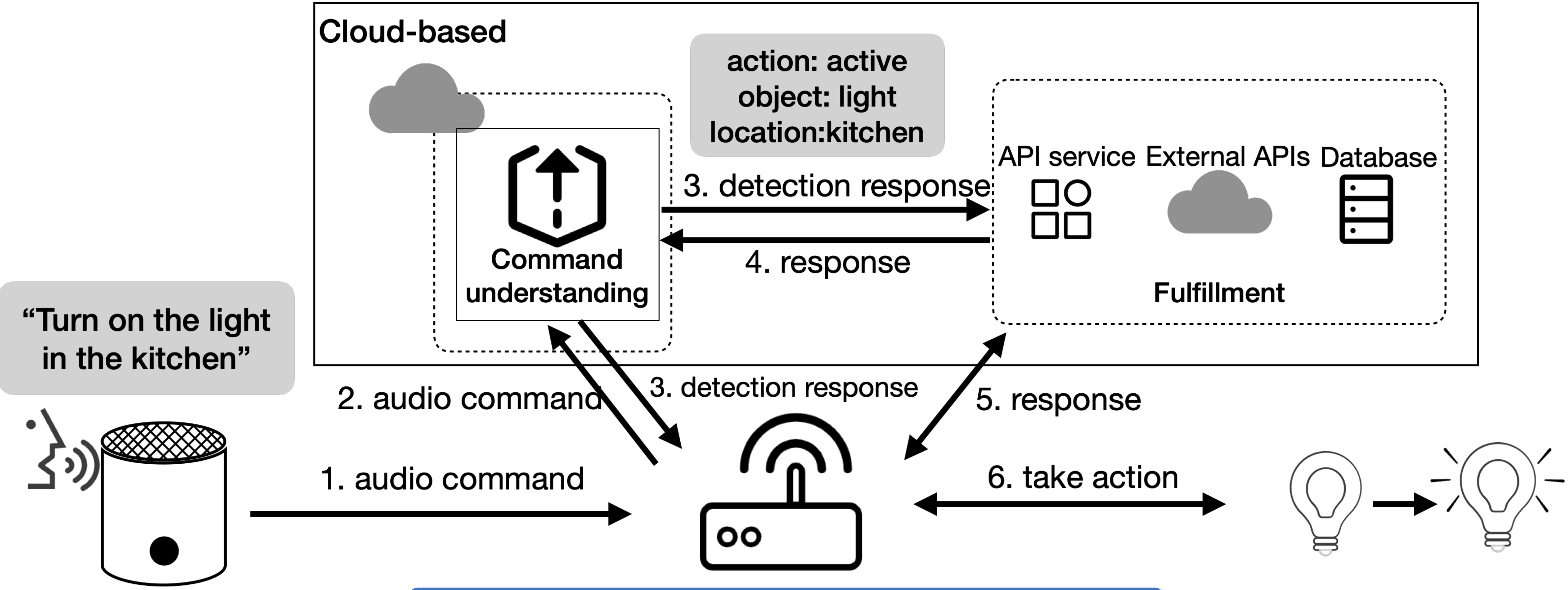
Source: Voicebot.ai 2020

## Q3 2019 market share (28.6 million)



S. Analytics, “Global smart speaker vendor & os shipment and installed base market share by region: Q4 2019,” 2020.

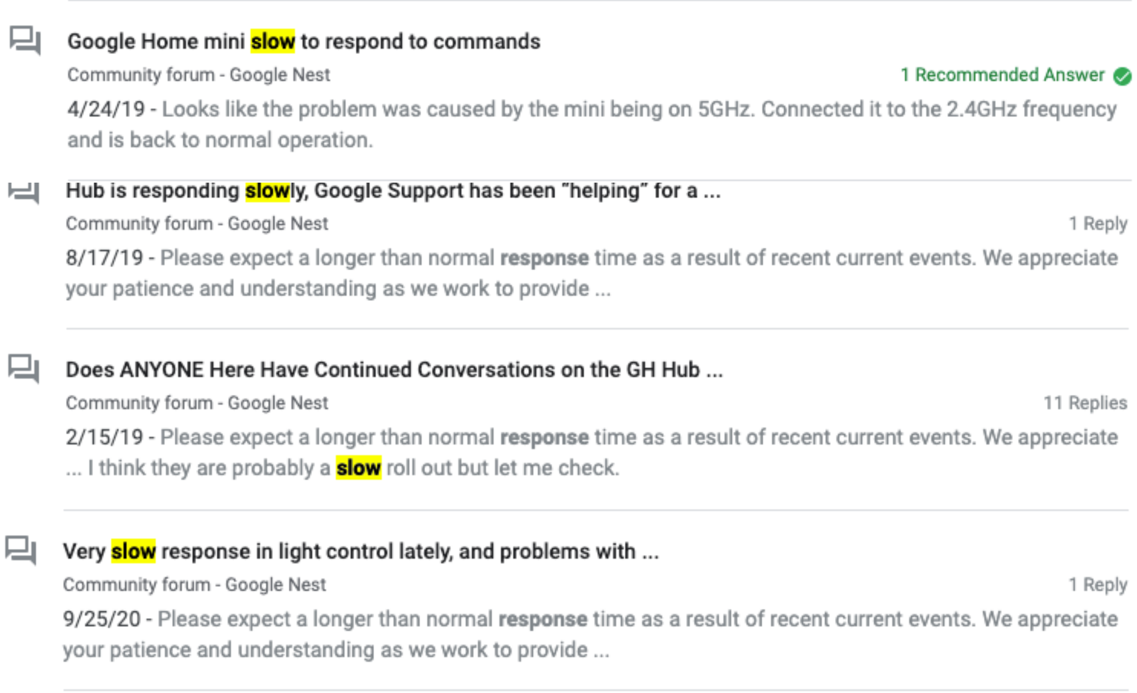
# Status-quo Approach



[Motivation 1] Command happens in home, fulfills in home.

# Limitations

- FAQ collected from Google and Amazon product forums

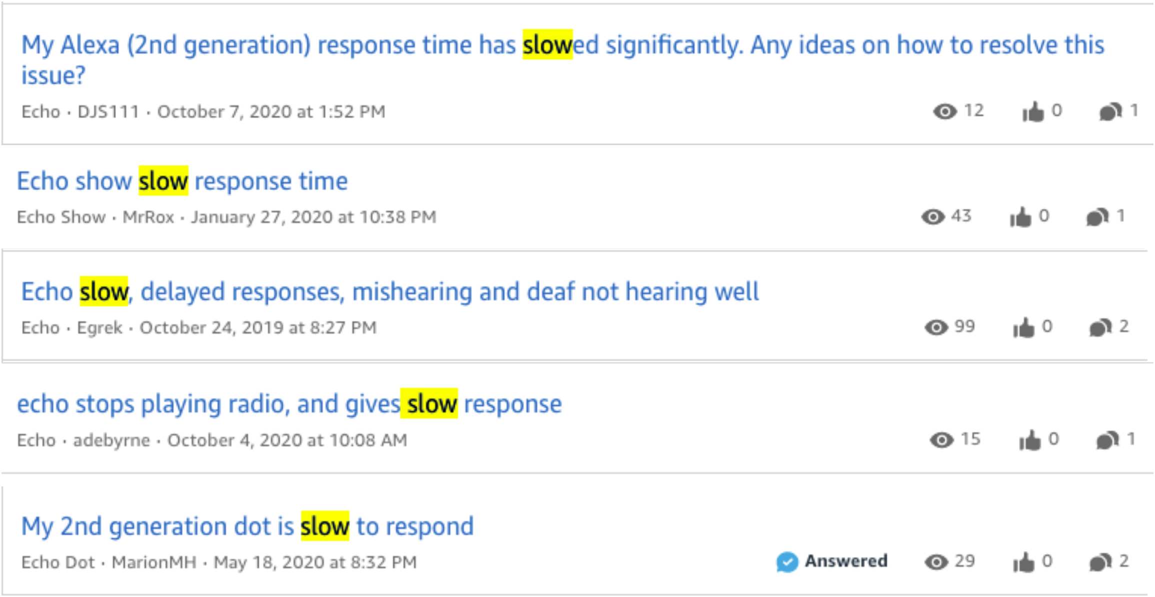


**Google Home mini slow to respond to commands**
Community forum - Google Nest — 1 Recommended Answer ✓
4/24/19 - Looks like the problem was caused by the mini being on 5GHz. Connected it to the 2.4GHz frequency and is back to normal operation.

**Hub is responding slowly, Google Support has been "helping" for a ...**
Community forum - Google Nest — 1 Reply
8/17/19 - Please expect a longer than normal **response** time as a result of recent current events. We appreciate your patience and understanding as we work to provide ...

**Does ANYONE Here Have Continued Conversations on the GH Hub ...**
Community forum - Google Nest — 11 Replies
2/15/19 - Please expect a longer than normal **response** time as a result of recent current events. We appreciate ... I think they are probably a **slow** roll out but let me check.

**Very slow response in light control lately, and problems with ...**
Community forum - Google Nest — 1 Reply
9/25/20 - Please expect a longer than normal **response** time as a result of recent current events. We appreciate your patience and understanding as we work to provide ...



My Alexa (2nd generation) response time has slowed significantly. Any ideas on how to resolve this issue?
Echo · DJS111 · October 7, 2020 at 1:52 PM — 👁 12 👍 0 💬 1

Echo show slow response time
Echo Show · MrRox · January 27, 2020 at 10:38 PM — 👁 43 👍 0 💬 1

Echo slow, delayed responses, mishearing and deaf not hearing well
Echo · Egrek · October 24, 2019 at 8:27 PM — 👁 99 👍 0 💬 2

echo stops playing radio, and gives slow response
Echo · adebyrne · October 4, 2020 at 10:08 AM — 👁 15 👍 0 💬 1

My 2nd generation dot is slow to respond
Echo Dot · MarionMH · May 18, 2020 at 8:32 PM — ✓ Answered 👁 29 👍 0 💬 2

[Motivation 2] Slow response, unstable performance harms user experience.

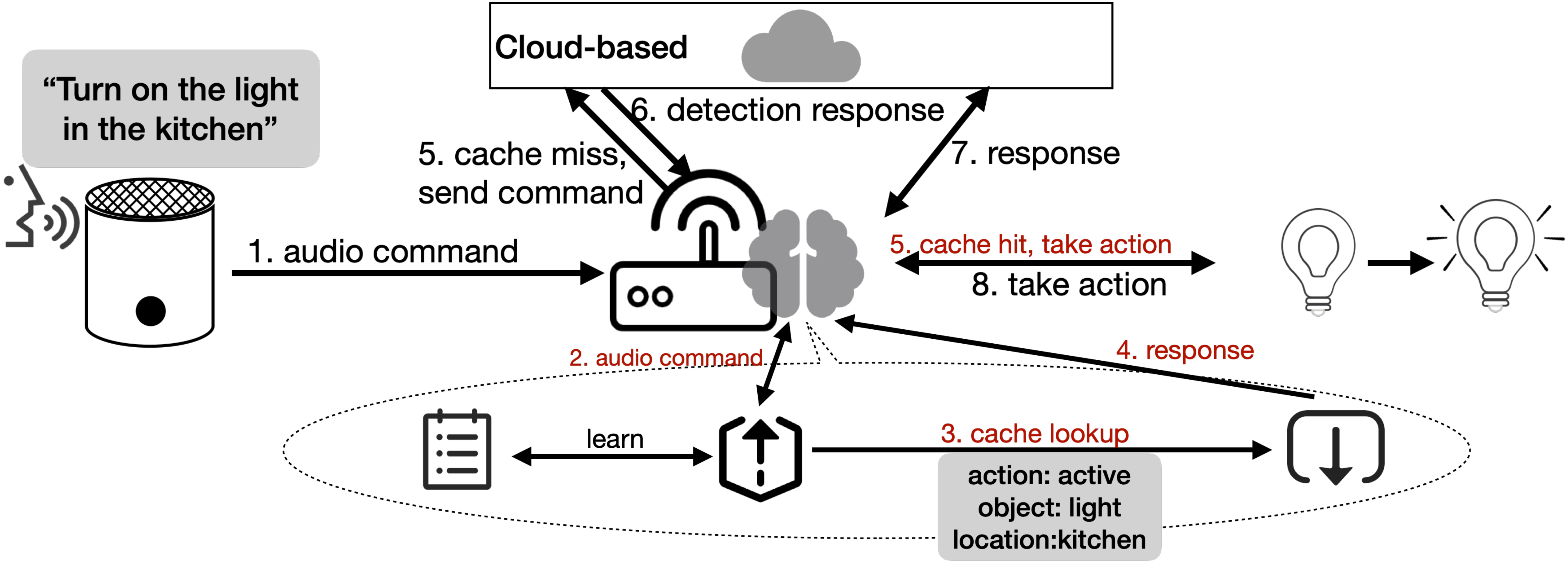
# User Behavior

- Google home usage survey[1]
  - 65,499 utterances, 88 diverse homes, over 110 days.
  - Limited command length: 1 – 10 words, median 4 words.
  - Highly spatial-temporality related:
    - ~ 3 domains/household.
    - Active usage 7AM – 11PM, peaks 5-6PM.
  - Semantic duplicated: frequently change commands for same information.

[Motivation 3] Smart home commands are short in length, limited in topic, and driven by intent

[1] F. Bentley, C. Luvogt, M. Silverman, R. Wirasinghe, B. White, and D. Lottridge, "Understanding the Long-Term Use of Smart Speaker Assistants," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, no. 3, pp. 1–24, Sep. 2018. [Online].

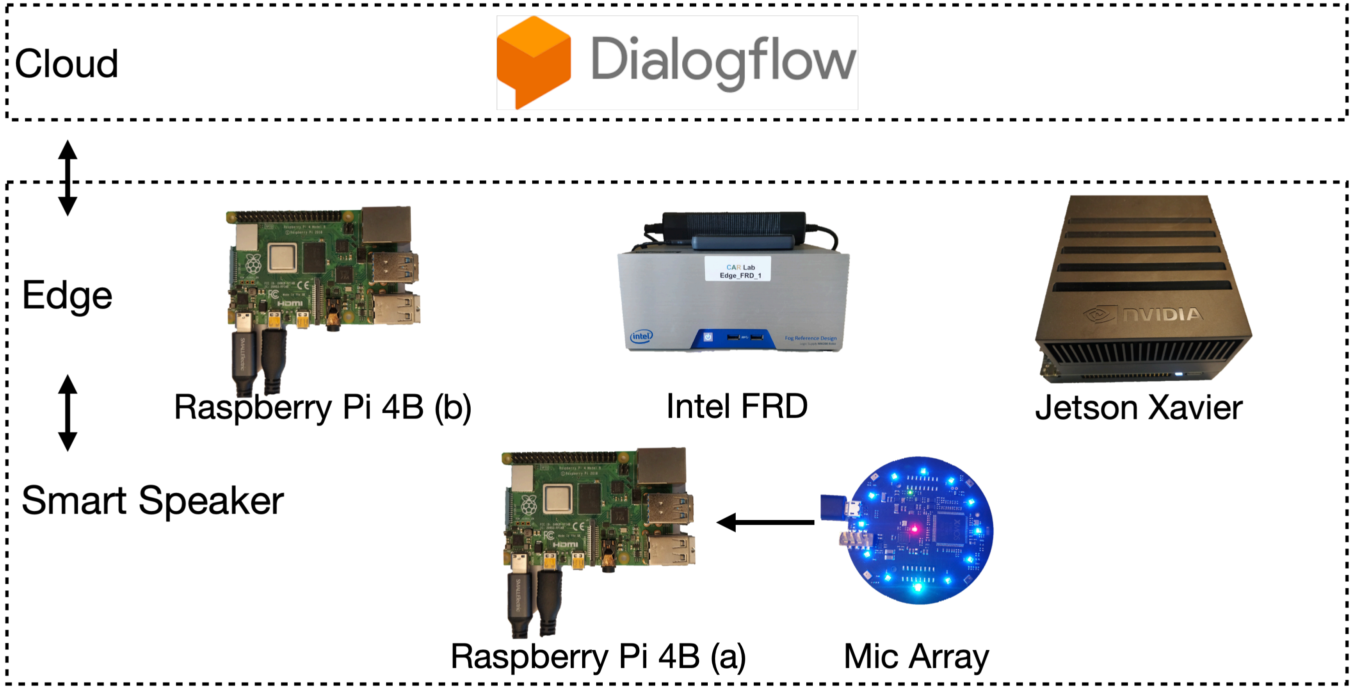
# CHA: An overview



# Contributions

- Identifying two drawbacks of the cloud-based voice assistant system.
- Developing an edge-based caching framework to improve user experience.
- Exploring system efficiency strategies for resource-constraint devices in home environment.

# Experiment Setup



Cloud

Dialogflow

Edge

Raspberry Pi 4B (b)

Intel FRD

Jetson Xavier

Smart Speaker

Raspberry Pi 4B (a)

Mic Array

| Hardware | CPU | GPU | Memory (GB) | Cost (USD) |
|---|---|---|---|---|
| Raspberry Pi 4B | ARMv7 | N/A | 4 | 55 |
| Intel Fog Reference Design | Intel Xeon E3-1275 | N/A | 32 | N/A |
| Jetson AGX Xavier | ARMv8 | 512-core Volta | 32 | 699 |

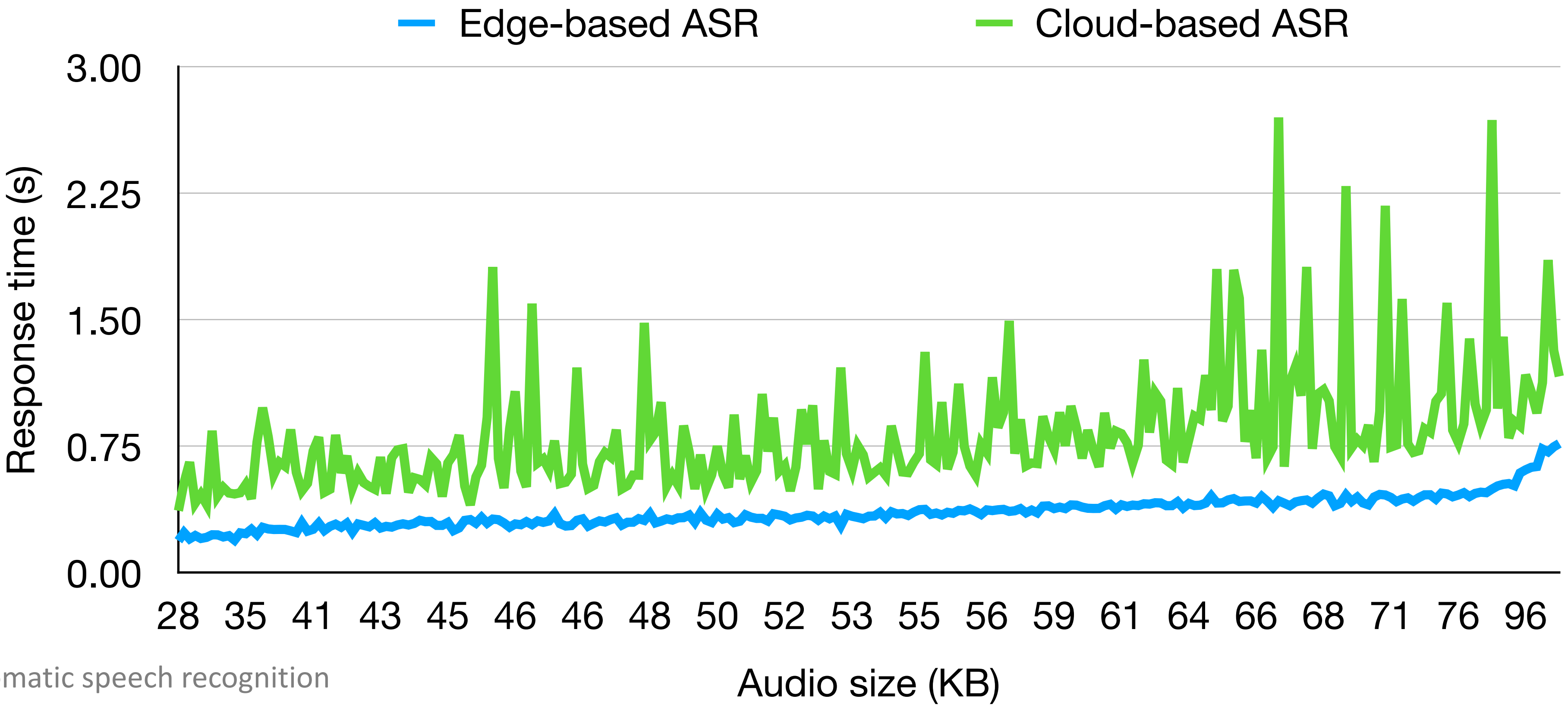
# Dataset

- Fluent Speech Commands
  - Typical smart home commands in English: home automation, task management.
  - 1 – 9 words / spoken command.
  - 31 intents, 3 slot types.
  - 4 – 24 types of expressions / intent. 248 unique utterances.

| Intent (trigger) | Commands |
|---|---|
| Increase volume | Louder please. |
| | Turn sound up. |
| | I can’t hear that. |
| | I need to hear this, increase the volume. |
| Active kitchen light | Turn on the kitchen light. |
| | Switch on the kitchen light. |
| | Kitchen light on. |

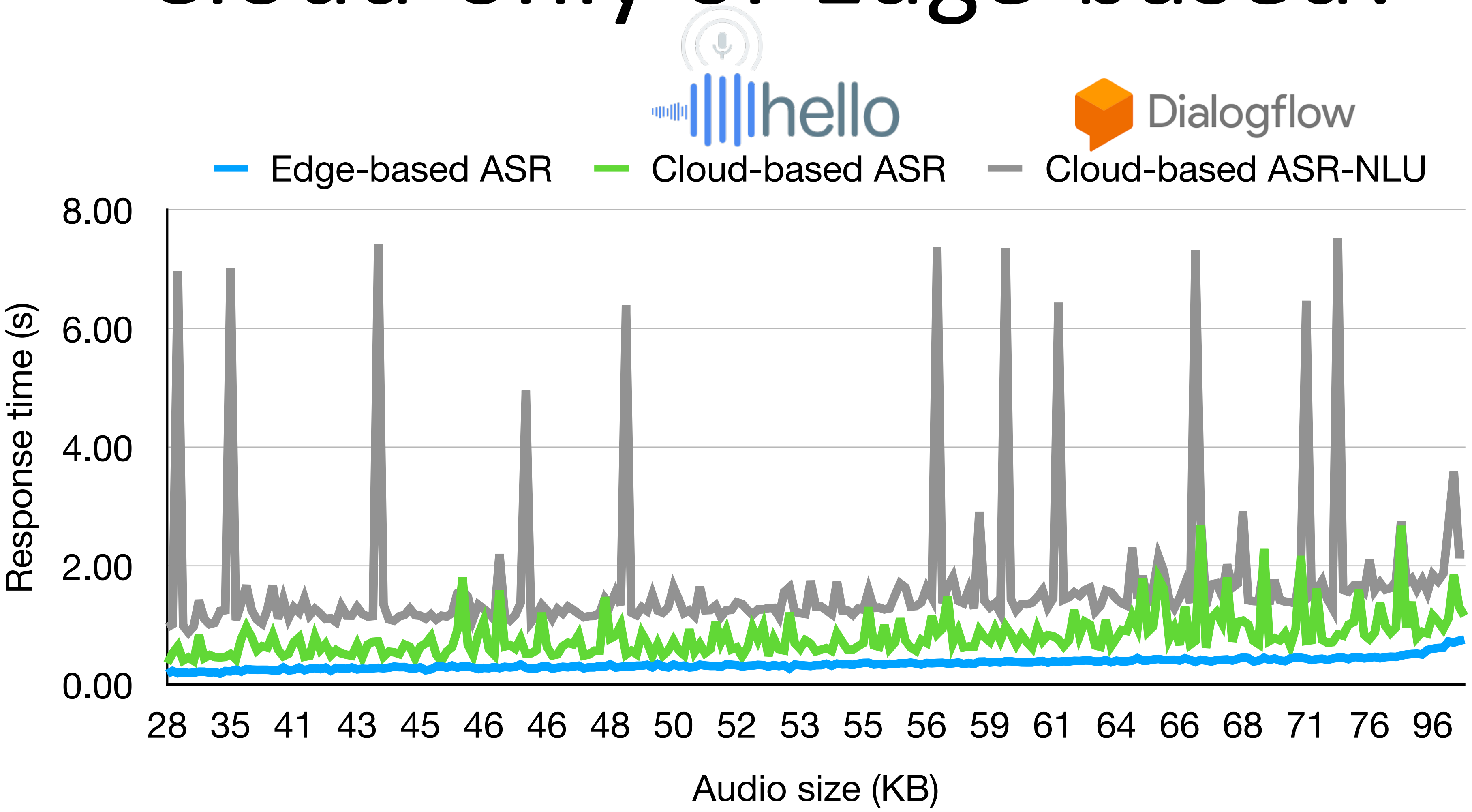
# Cloud-only or Edge-based?

| | Word error rate (WER) | Sentence accuracy |
|---|---|---|
| Cloud-only ASR | 10.42% | 83.19% |
| Edge-based ASR | 2.52% | 96.12% |



ASR: Automatic speech recognition

# Cloud-only or Edge-based?



Edge brings lower latency, more stable performance comparing to cloud-only processing.

NLU: Nat

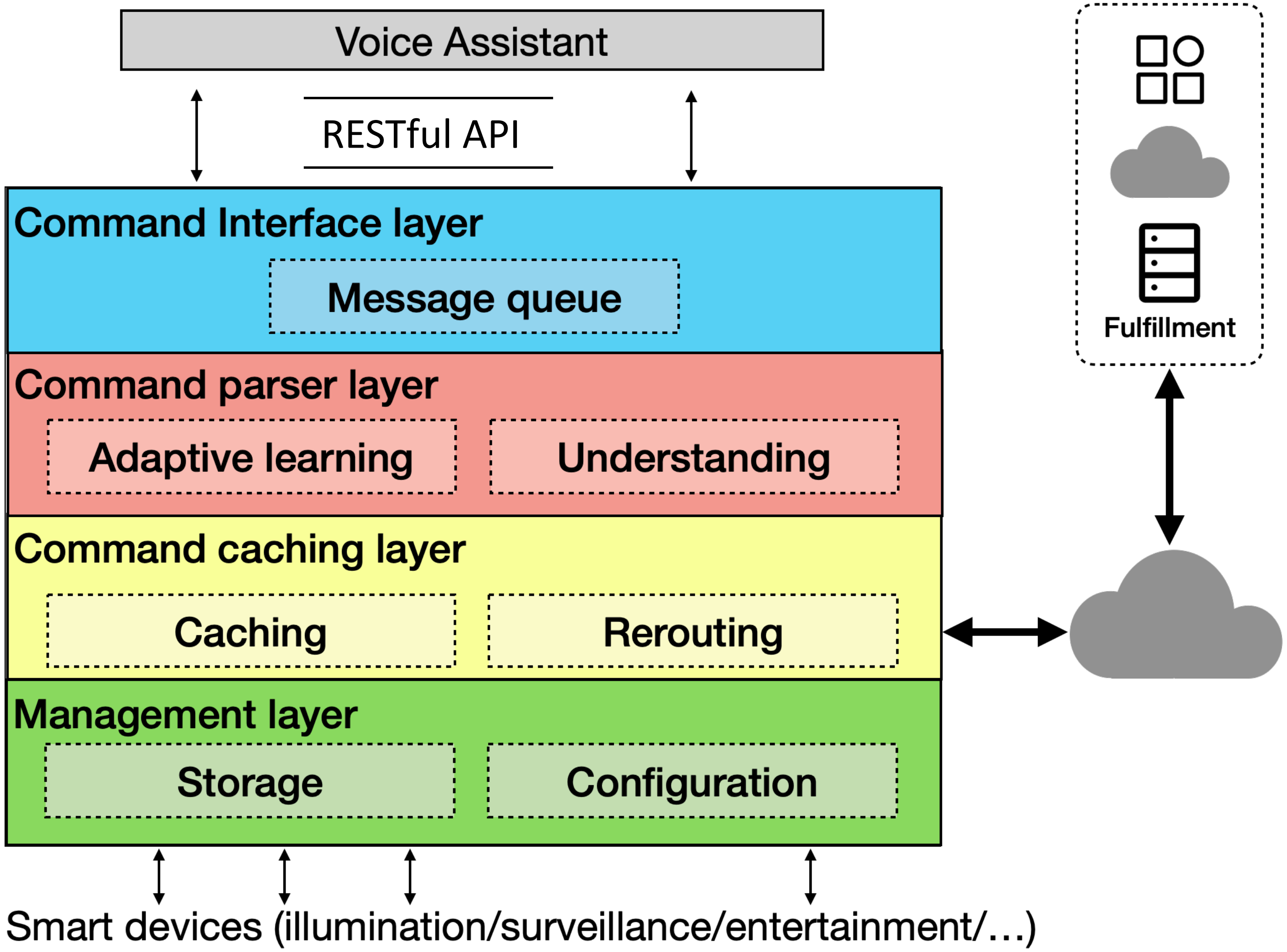
# System Design

Response latency
Understanding accuracy
System efficiency

“Turn on the light in the kitchen”
→
Intent (trigger): active_kitchen_light

Hash table
<key: trigger, value: action>

Trigger: “active kitchen light”
Entity: light.kitchen
Status: (state == off)
Action: state.on



# Command Understanding

| | Turn | On | The | Light | In | The | kitchen |
|---|---|---|---|---|---|---|---|
| Slot | B-active | I-active | O | B-object | O | O | B-location |
| Intent | Active_kitchen_light | | | | | | |

- Goal
  - Audio input → (intent, slot)
- Methodology
  - Automatic speech recognition + natural language understanding (ASR + NLU)
    - Conventional method
  - Spoken language understanding (SLU)
    - Extracts words and phoneme features
    - followed by intent detection and slot filling
- CHA
  - ASR: pocketsphinx[2]
  - NLU: BERT[3]

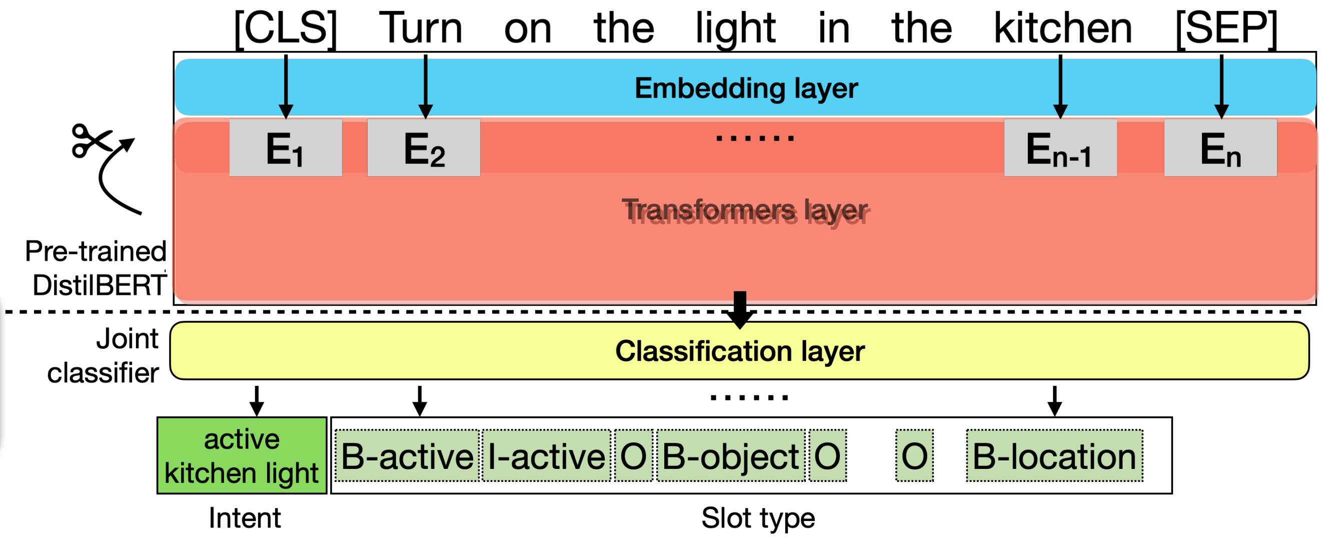
[2] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 1. IEEE, 2006, pp. I–I.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805 [cs], May 2019, arXiv: 1810.04805.

# Command Understanding (cont’d)

- Inherit from BERT
  - Pre-trained distilBERT
  - Jointly detect intent and slot types

Improve for cache miss? Pruning layers


[CLS] Turn on the light in the kitchen [SEP]

Embedding layer

$E_1$ $E_2$ …… $E_{n-1}$ $E_n$

Transformers layer

Pre-trained DistilBERT

Joint classifier

Classification layer

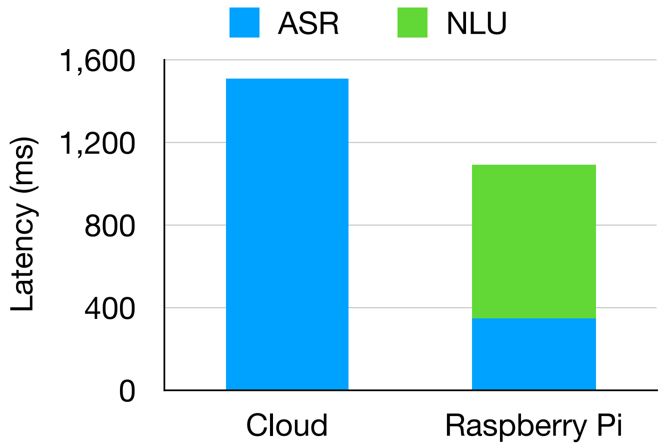
……

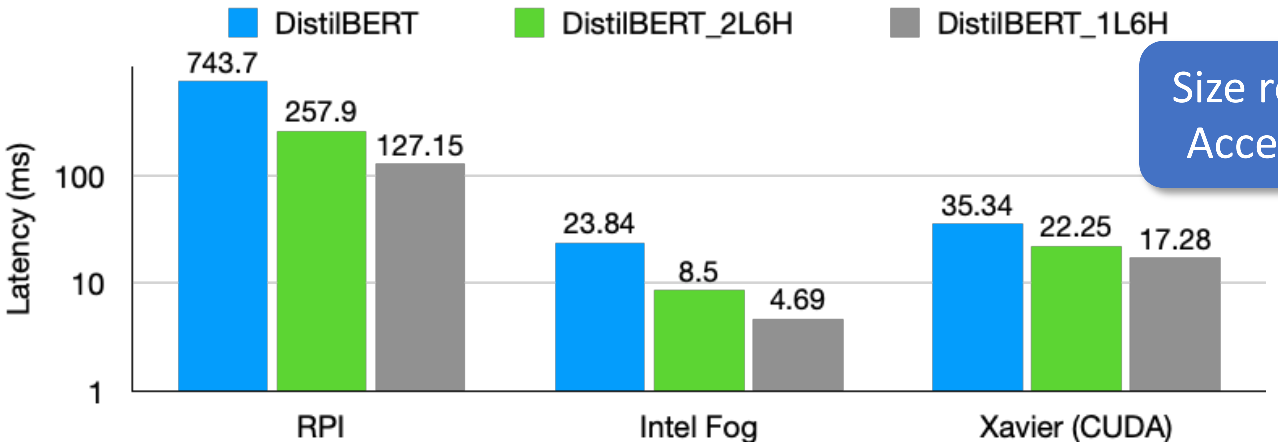
active kitchen light

B-active I-active O B-object O O B-location

Intent

Slot type

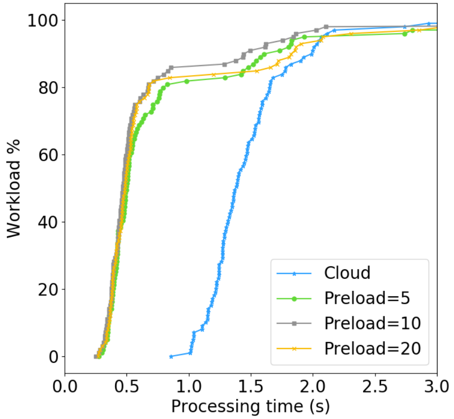




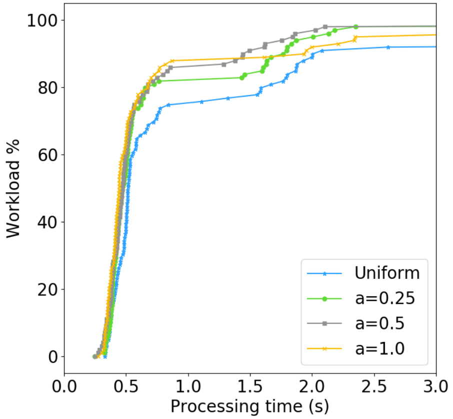
Size reduction: 53%
Acceleration: 5.8X

# System Efficiency

- Workload
  - Simulate query in Pareto distribution.
    - Probability distribution $f(trigger, \alpha) = \frac{\alpha}{trigger^{\alpha+1}}$. Higher $\alpha$ has higher semantic locality.
    - $\alpha$ = 0.25, 0.5, 1.0, and uniform distribution.
    - Cache warmup with 5, 10, 20 commands.



$\alpha$ = 0.5

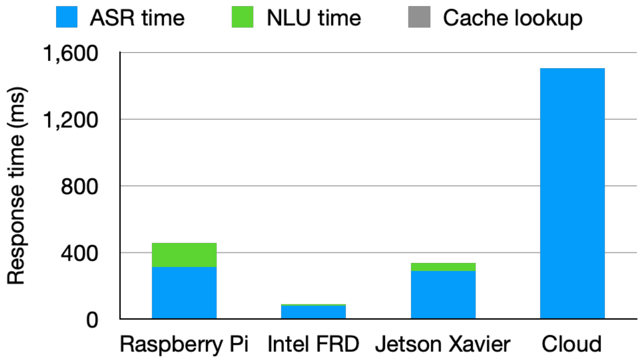


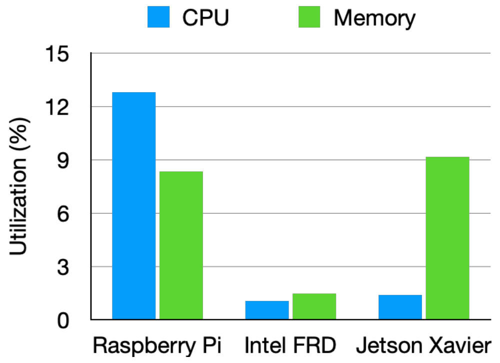
Warmup with 10 commands

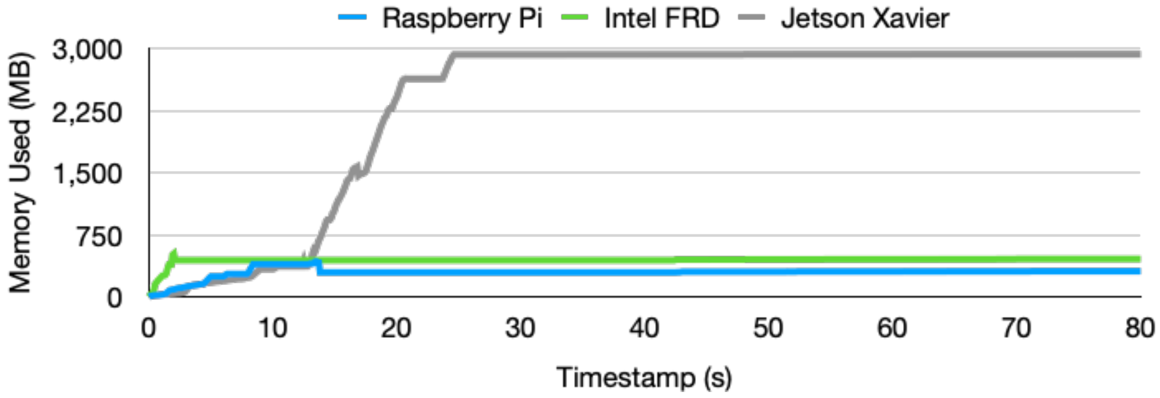
- Insight
  - On Raspberry Pi, CHA provides a fast and stable response with a lightweight understanding module.

# CHA on Different Edge Devices

- Response time
  - Reduced by 70%, 94%, 77% than the cloud-only solution for cache hit item.
  - Low overhead for cache missed item.
- Resource utilization
  - Low resource consumption across platforms.
  - System loading takes 13, 2, 24 seconds on three platforms, respectively.
- CHA has generality to be deployed on different hardware equipped devices.







# Discussion

- Layer pruning benefits BERT and its variants with subtle performance degradation (when pruned to 1 layer).

| | Layers | Model size (MB) | Param size (million) | Intent accuracy | Slot F1 score |
|---|---|---|---|---|---|
| BERT | 12 → 1 | 438 → 126 | 110 → 30 | 96% → 92% | 96.3% |
| DistilBERT | 6 → 1 | 256 → 123 | 66 → 30 | 92% | 96.3% |
| ALBERT | 1 | 46.87 | 12 | 96% | 96.3% |

- End-to-end SLU model compression is challenging due to is dense and informative structure (compare to compressed NLU model).

| | Raspberry Pi | Intel FRD | Jetson Xavier |
|---|---|---|---|
| Inference time | 737.0 ms (127.2 ms) | 41.4 ms | 83.0 ms |
| Model size | 15.9 MB (123.8 MB) | | |
| Parameter size | 3 million (30 million) | | |

# Conclusion and Future Work

- Conclusion
  - CHA is proposed to address two drawbacks for cloud-based voice assistant systems.
  - CHA integrates a set of compression strategies to provide affordable and practical solution for home-based voice assistant systems.
  - CHA provides a 70% acceleration in voice command processing on the low-cost, resource-constrained raspberry pi, with low resource consumption.

- Future work
  - Exploring audio caching.
  - Developing model compression strategies.

# Thank you!

http://thecarlab.org/

xu.lanyu@wayne.edu