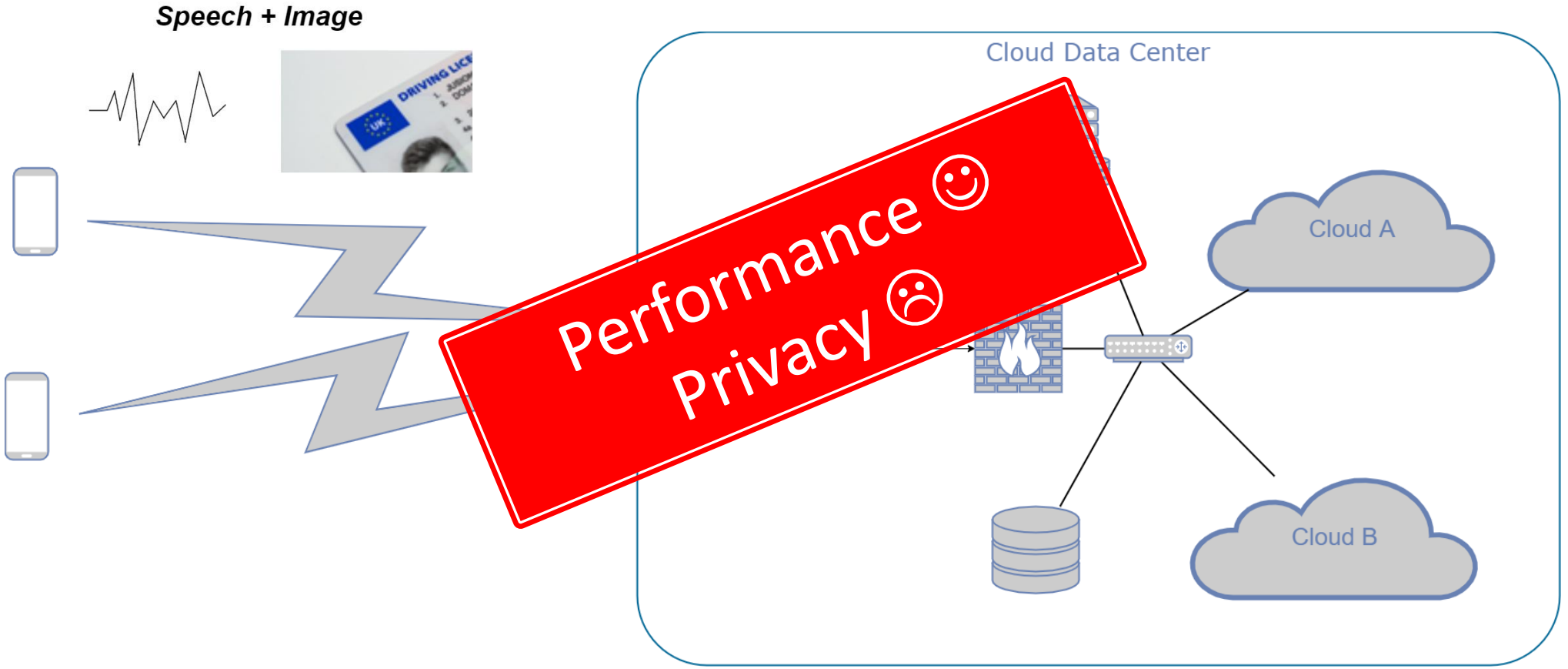
Privacy Preserving Intelligent Personal Assistant at the EdGE (PAIGE)

Yilei Liang (King's College London)

Dan O'Keeffe (Royal Holloway University of London)

Nishanth Sastry (King's College London)

Intelligent Personal Assistant (IPA) workload



Data leak cases

Apple apologises for allowing workers to listen to Siri recordings

r a

Contractors graded accidental activations including recordings of users having sex

▲ Apple has apologised to Siri users for not 'fully living up to our ideals'. Photograph: Bloomberg/Getty

Apple has apologised for allowing contractors to listen to voice recordings of Siri users in order to grade them.

The company made the announcement after it completed a review of the grading programme, which had been triggered by a Guardian report revealing its existence.

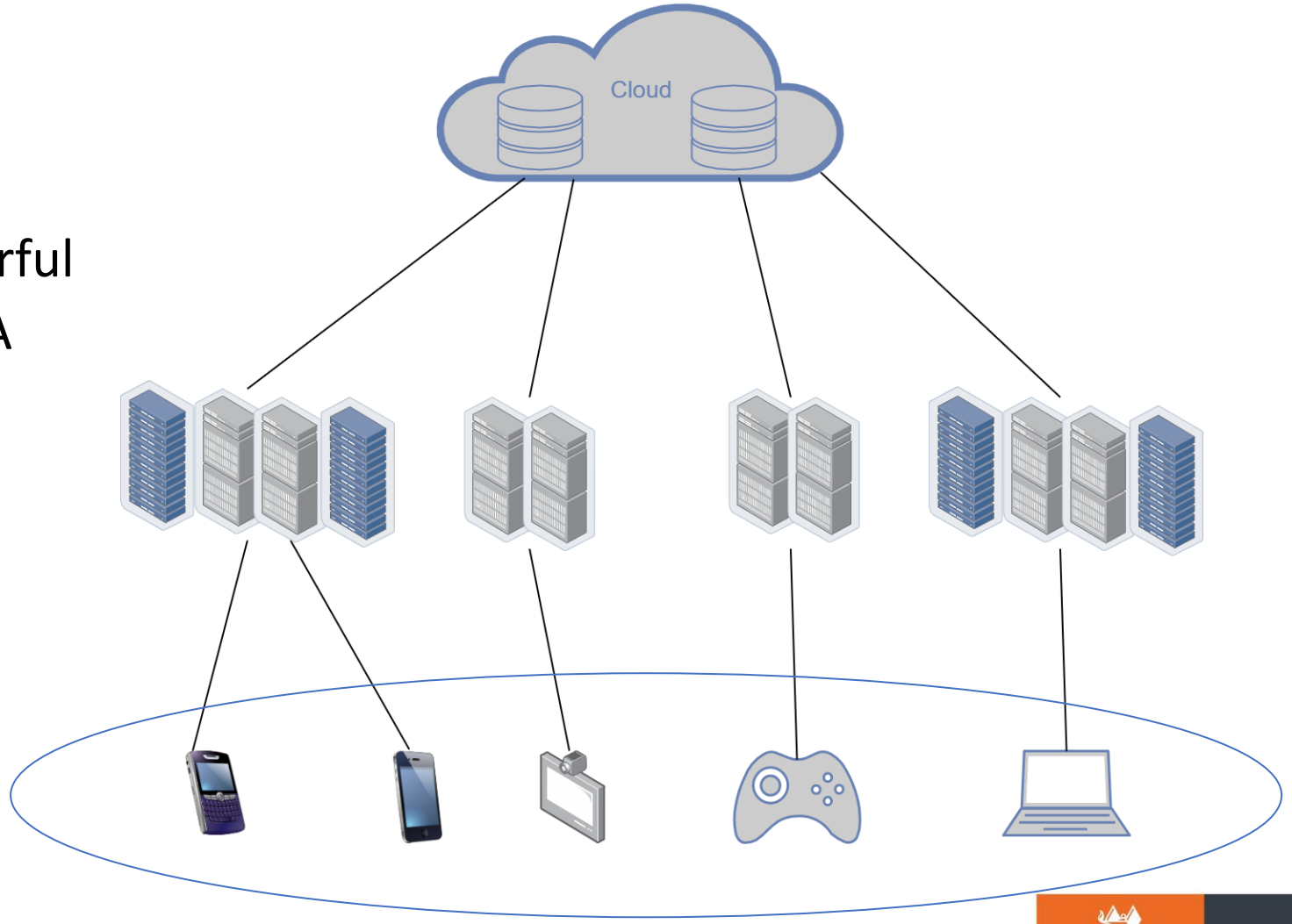
According to multiple former graders, accidental activations were regularly sent for review, having recorded confidential information, illegal acts, and even Siri users having sex.

"As a result of our review, we realise we have not been fully living up to our high ideals, and for that we apologise," Apple said in an unsigned statement posted to its website. "As we previously announced, we halted the Siri

Is Edge a solution?

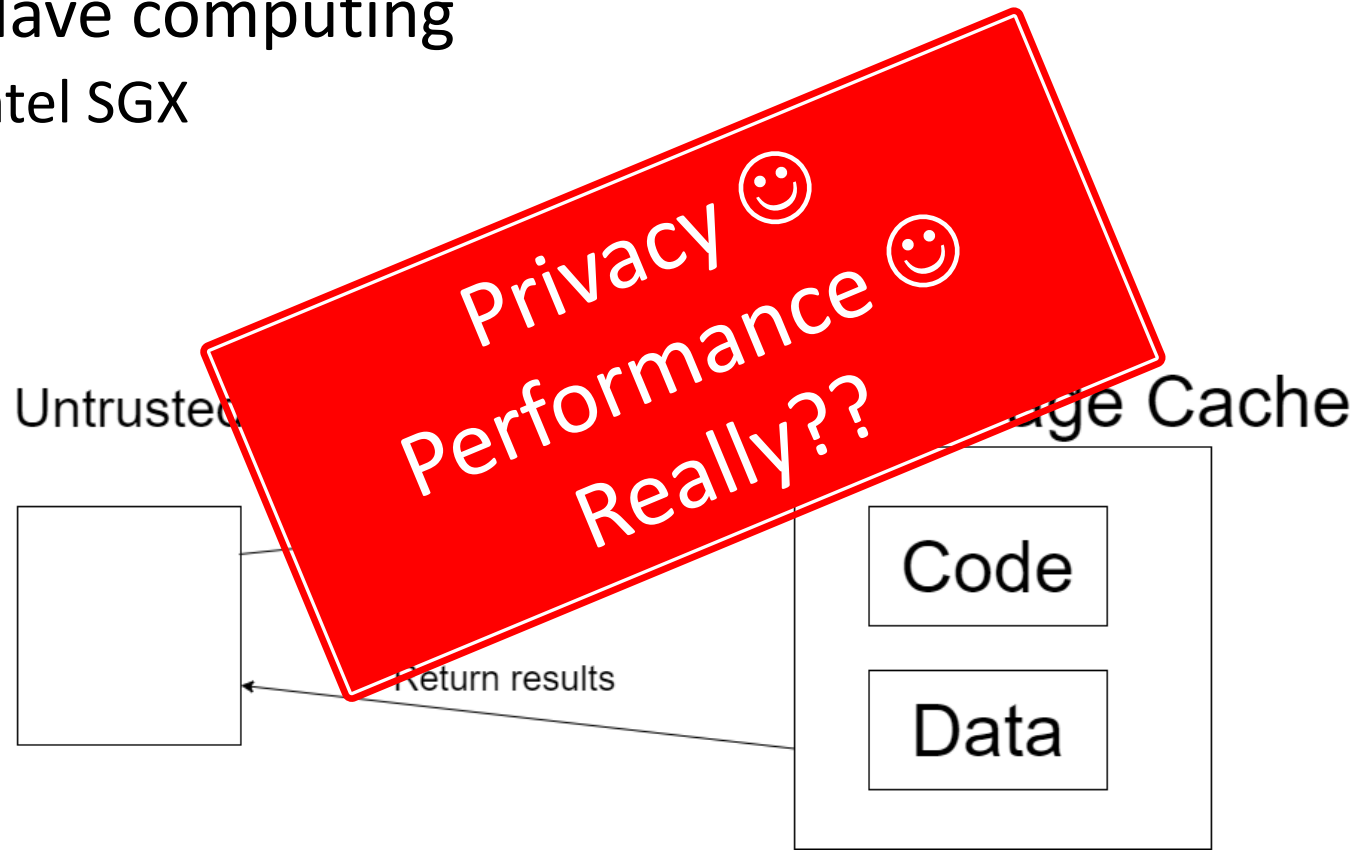
User edge devices are not powerful

Require a large database for Q/A



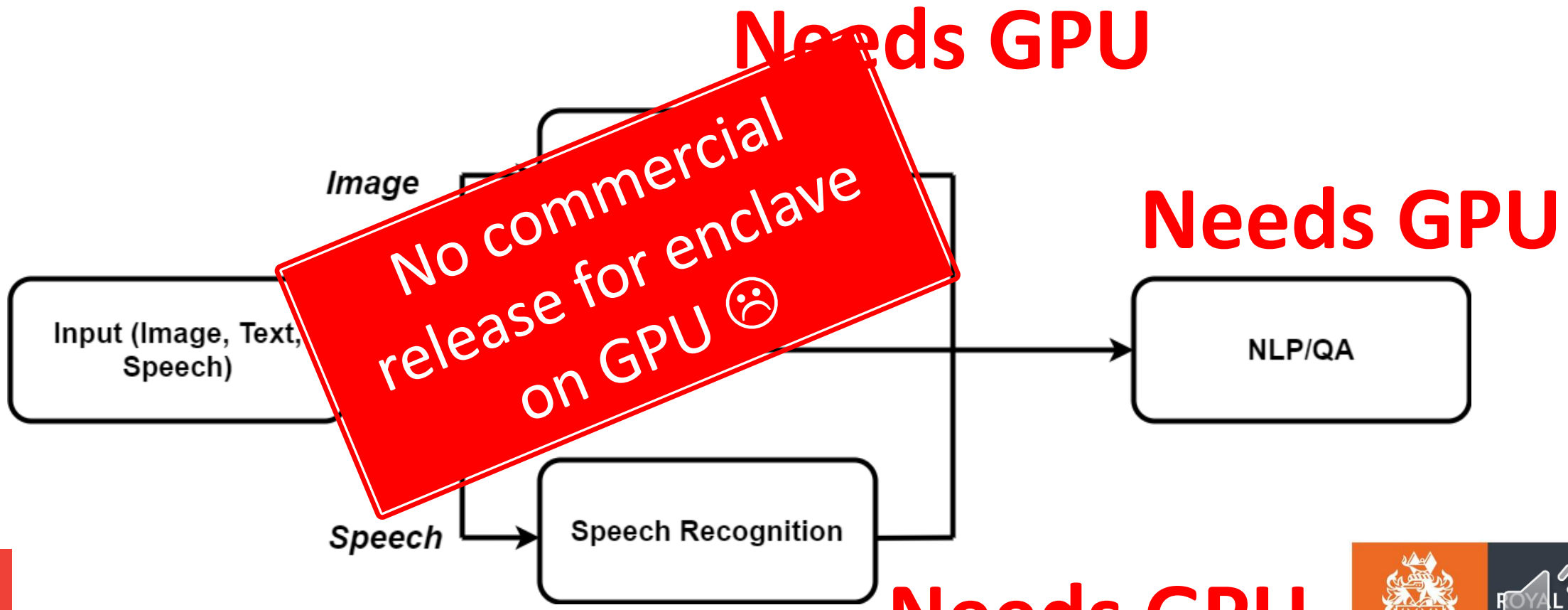
Can we preserve privacy in the cloud?

- Yes, enclave computing
  - E.g. Intel SGX



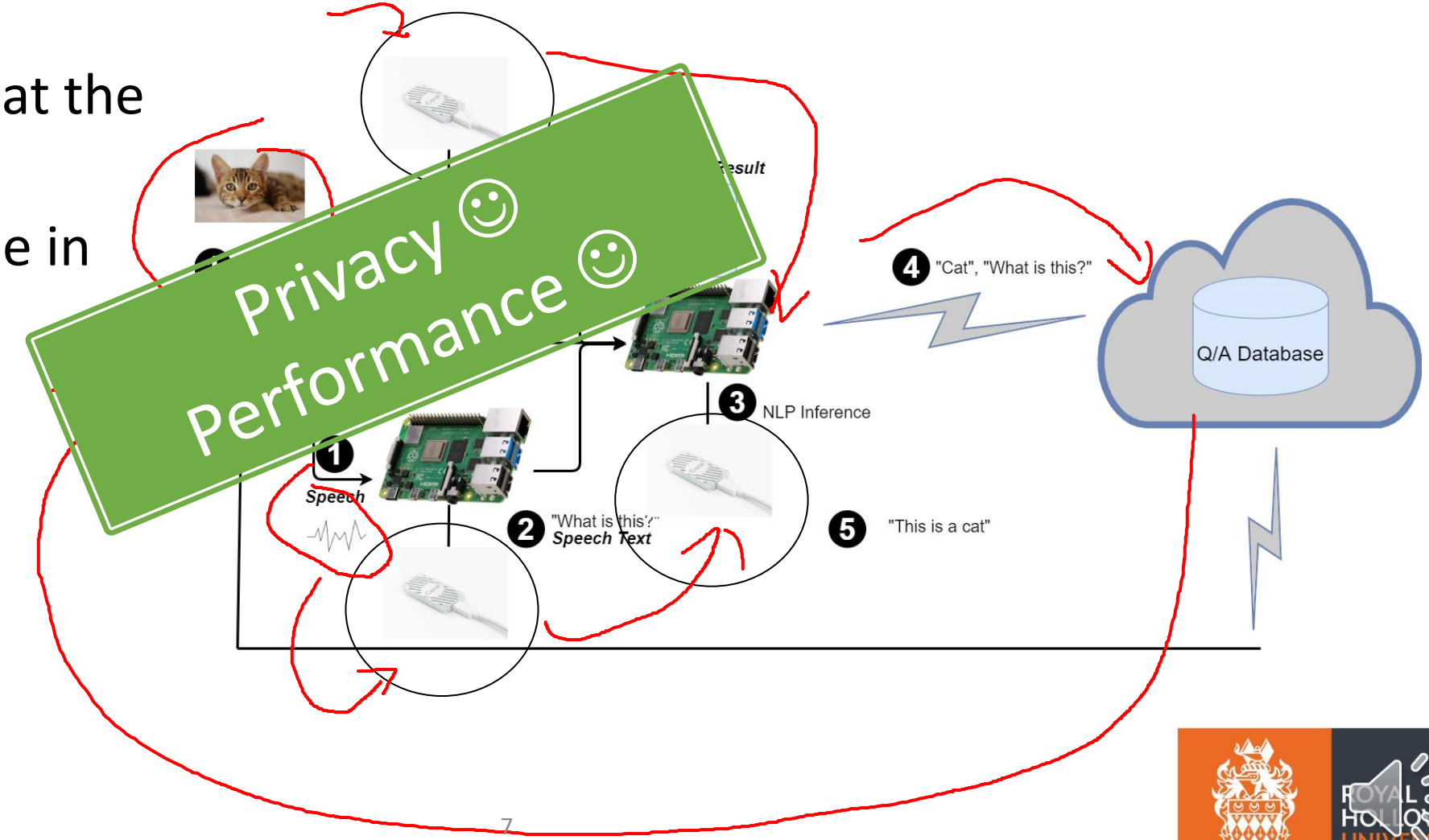
Intelligent Personal Assistant (IPA) workload

- Private Intelligence Assistant



Our solution – Hybrid Privacy Preserving IPA at the edge (PAIGE)

- Add accelerators at the Edge
- Keep the database in the cloud



Evaluation Goals

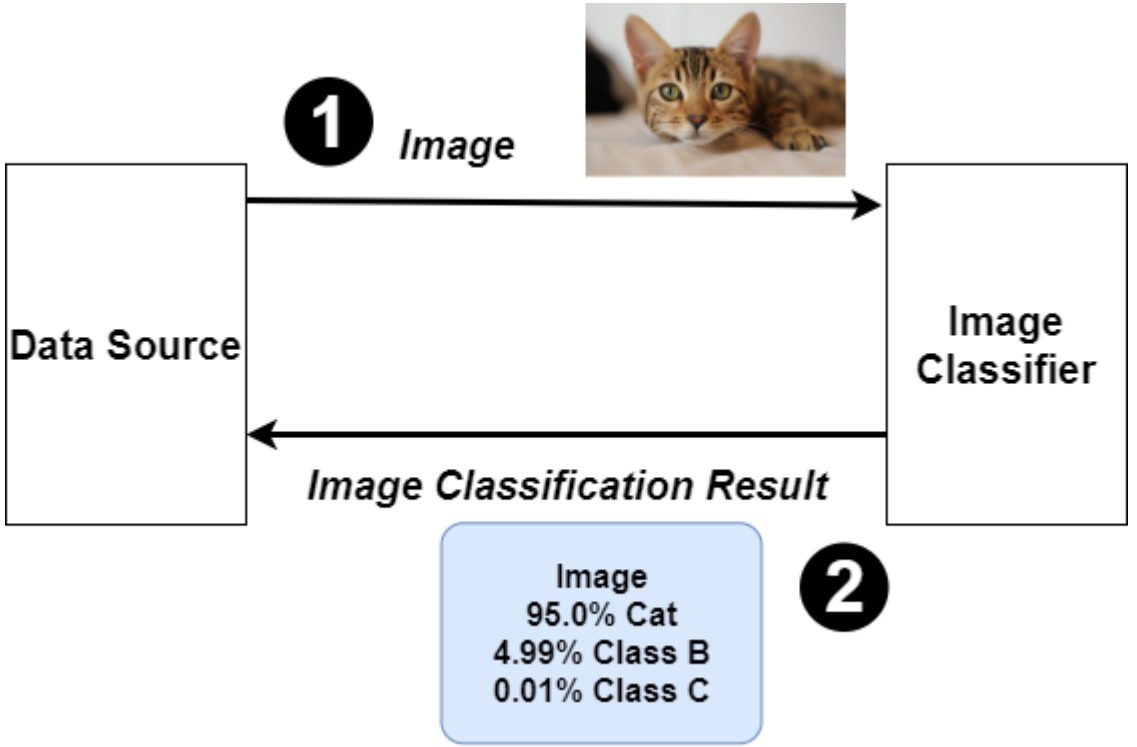
- Workload
  - Focus on image recognition
  - Future Work: Speech recognition, Question-Answering, NLP…

- What we measure
  - ML Performance at the Edge
  - Energy Consumption of Edge Devices

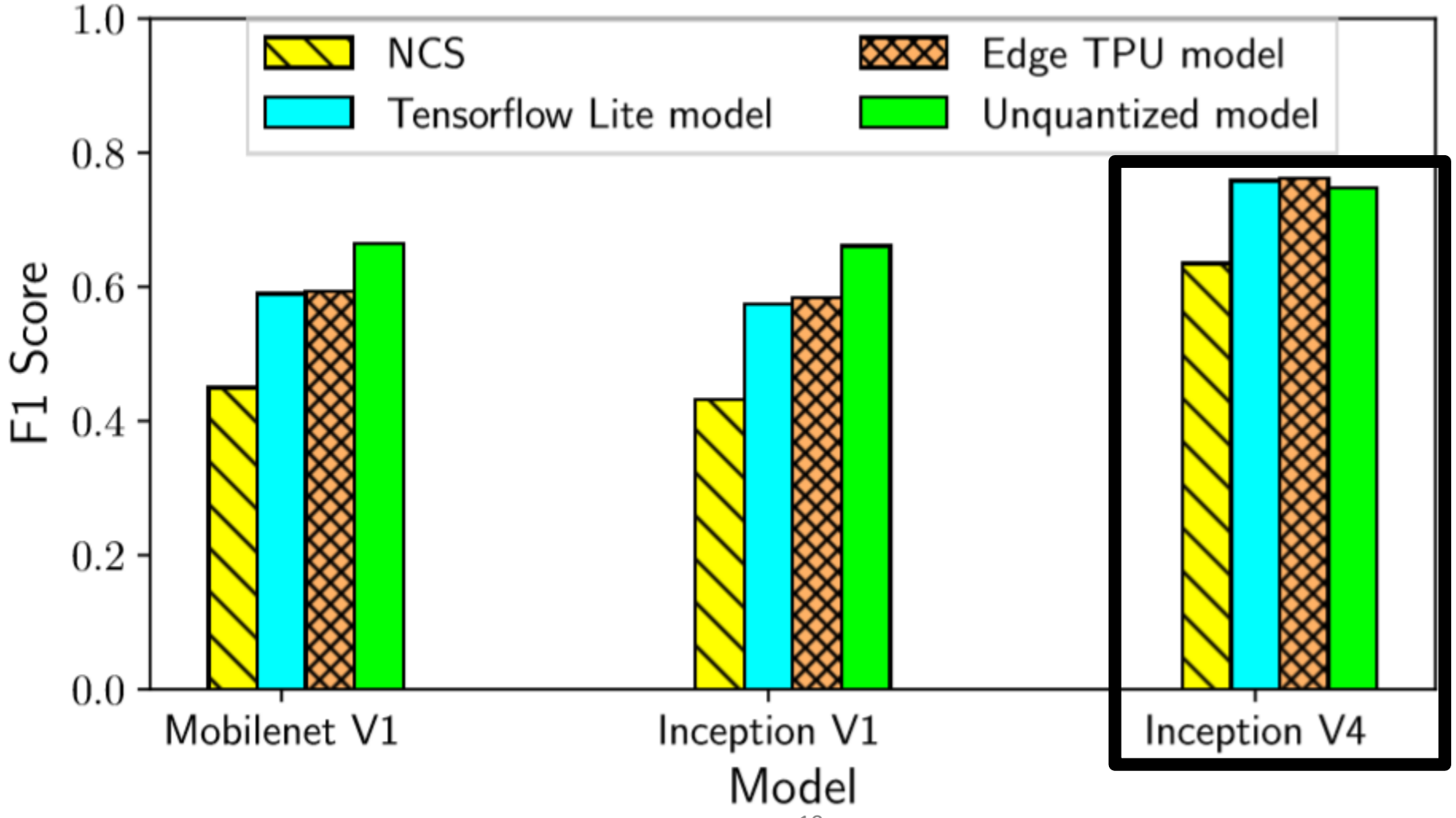
Across heterogeneity of devices and ML architectures

Evaluation on Image Recognition

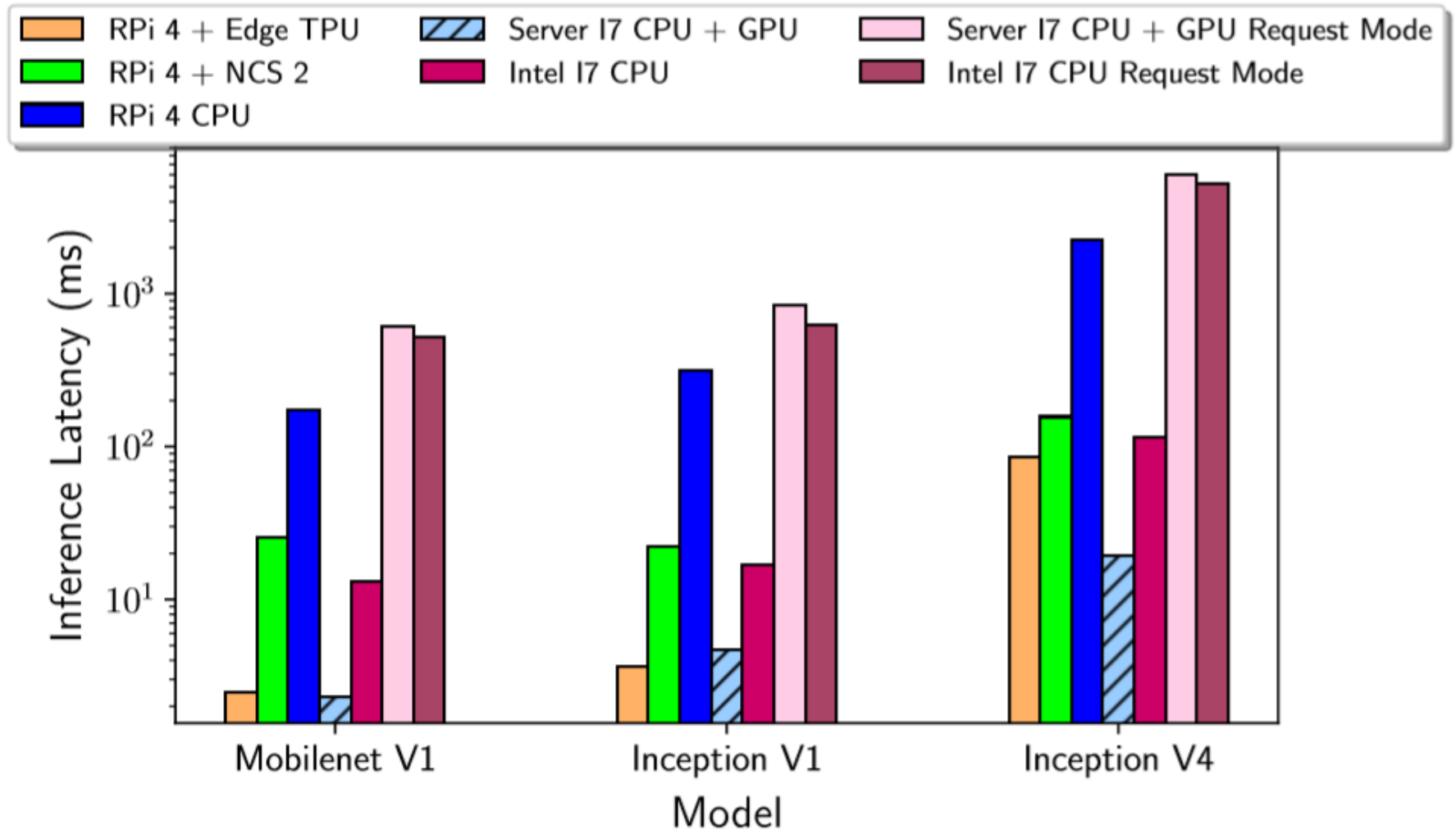
- Hardware Architecture
  - Raspberry Pi 4 (4GB RAM)
    - **RPi 4 CPU**
    - Neural Compute Stick 1st & **2nd Gen** (NCS 2)
    - **EdgeTPU**
  - Server Class CPU (E5645, **I7 8750H**)
  - **GPU (Nvidia RTX 2080 MAX-Q Design)**
- ML Architecture
  - Mobilenet **V1**, V2
  - Inception **V1**, V2, V3, **V4**



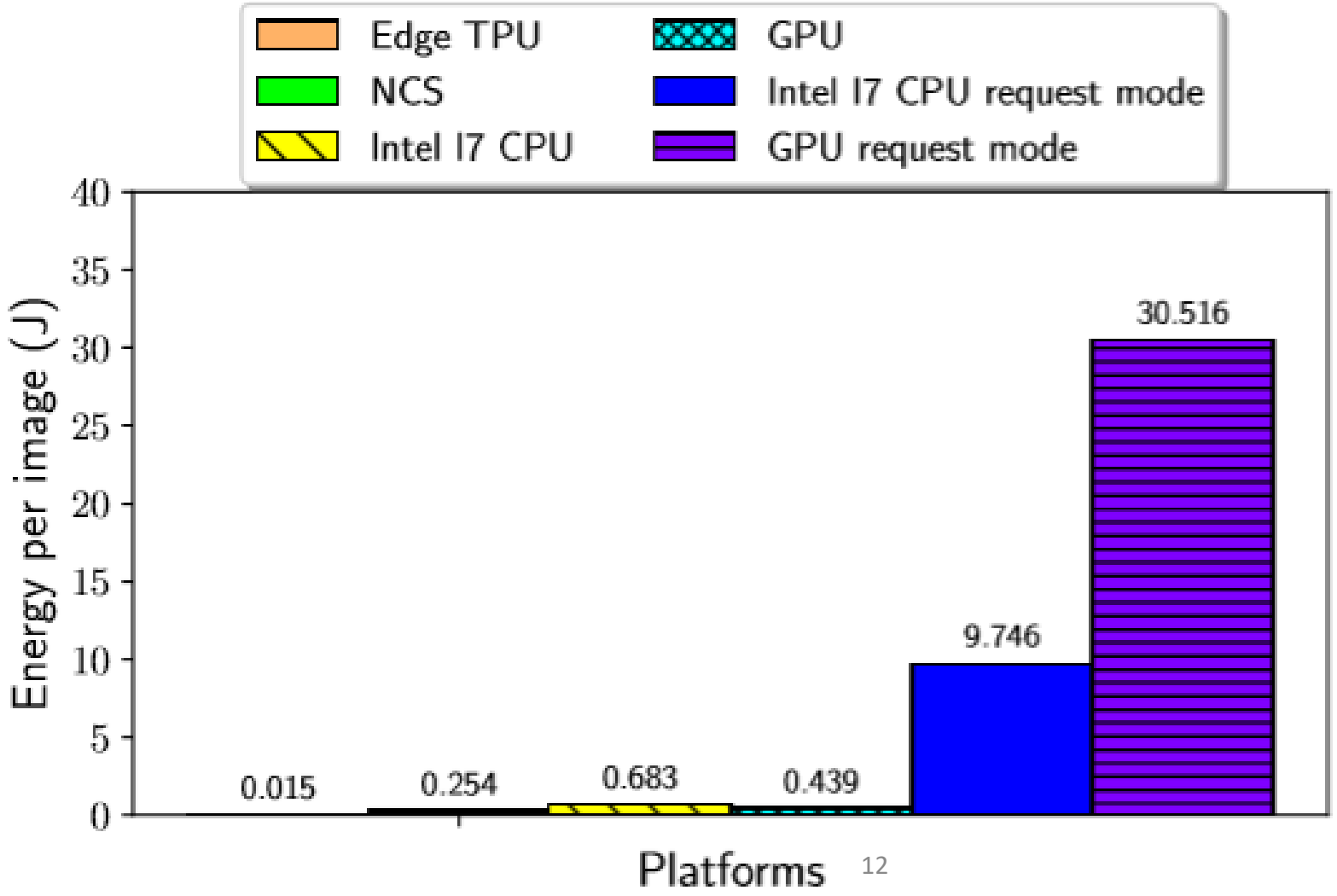
ML Performance Benchmark (F1 Score)



Inference Time Benchmark

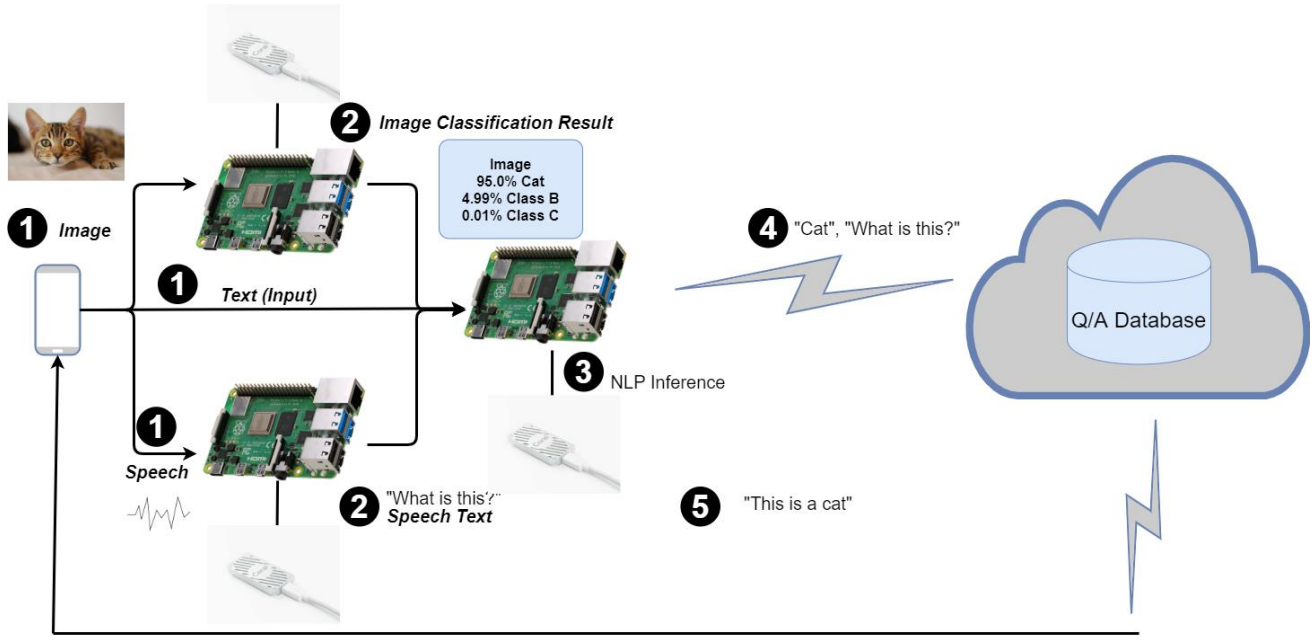


Energy Consumption Benchmark



Takeaways

- RPi + Edge accelerators have:
  - Similar performance to servers + GPU
  - Significantly lower energy consumption
- GPU still wins for larger models.



- Yilei.liang@kcl.ac.uk