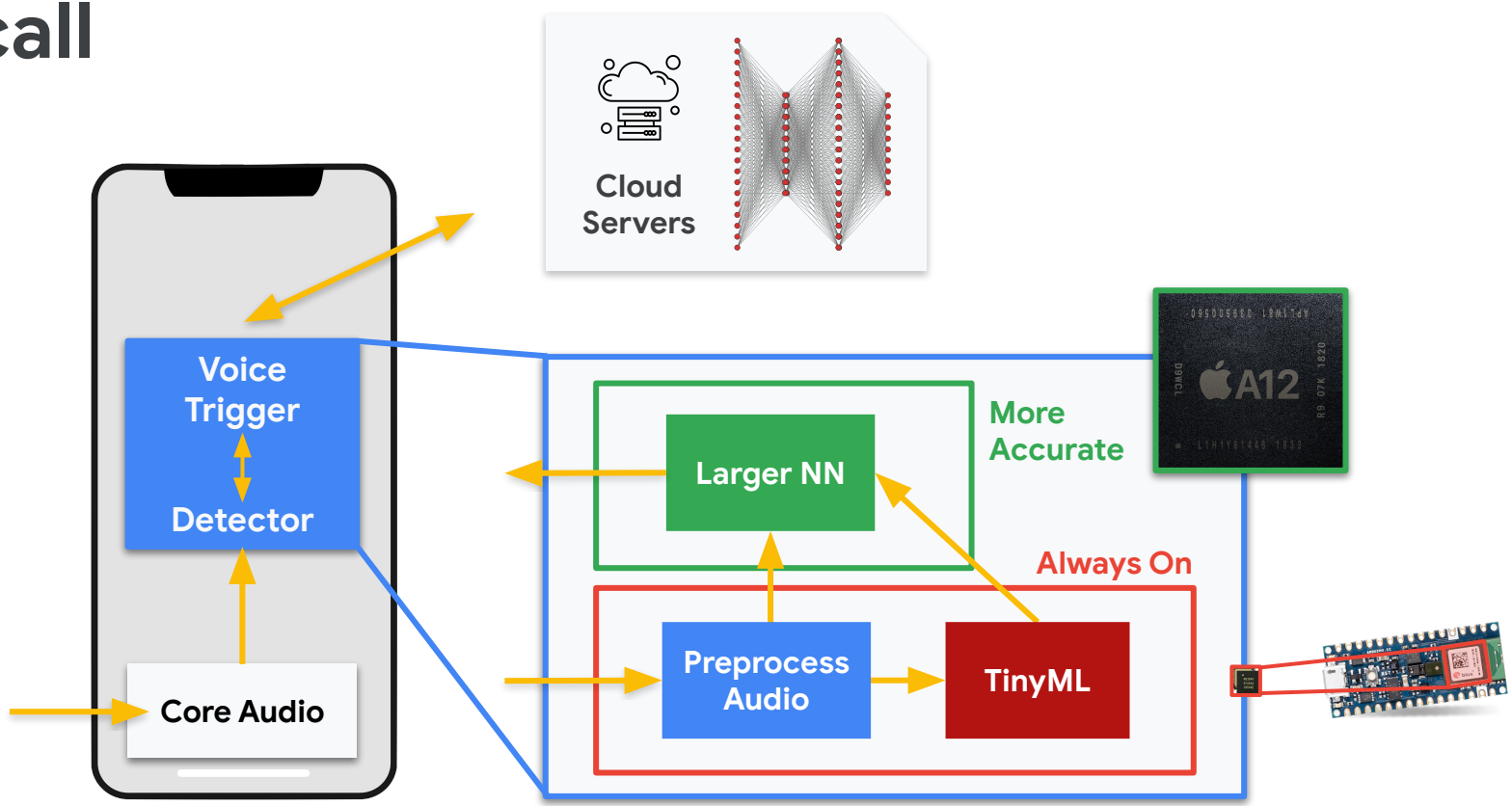
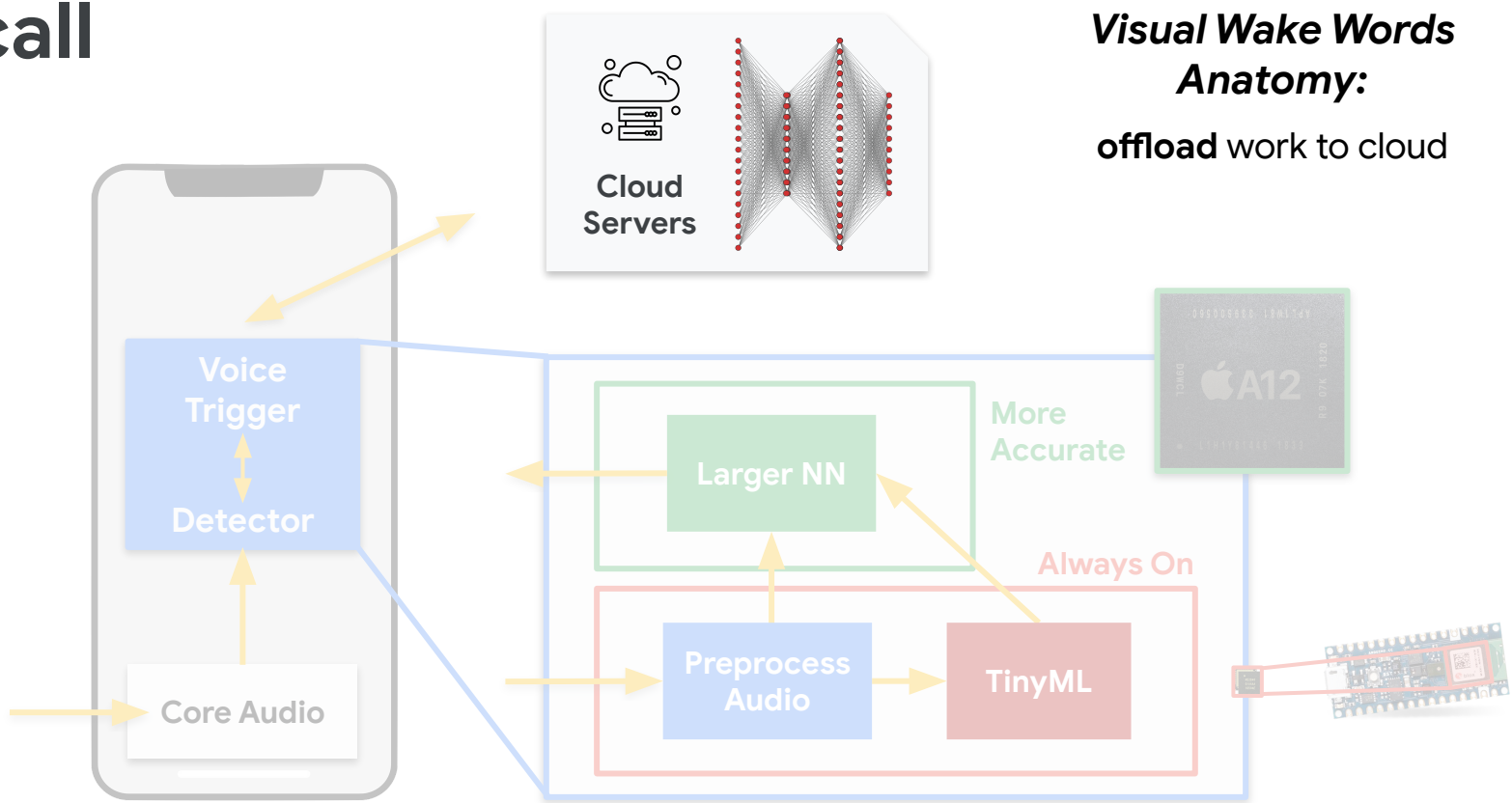
Visual Wake Words Challenges

Recall


Cloud
Servers
Voice
Trigger
Detector
Core Audio
Larger NN
More
Accurate
A12
Always On
Preprocess
Audio
TinyML

Recall



Visual Wake Words Anatomy:

offload work to cloud

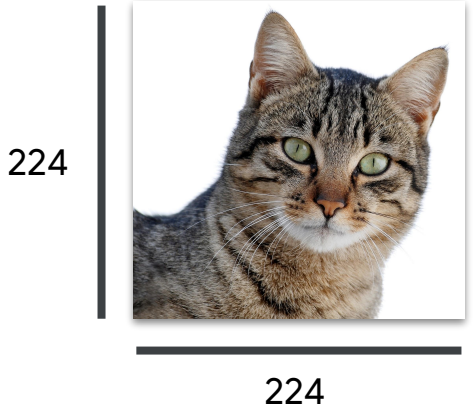
Simple **Experiment**

$$224 \times 224 \times 3 \times 4 = 602{,}112 \text{ Bytes}$$

Pixels

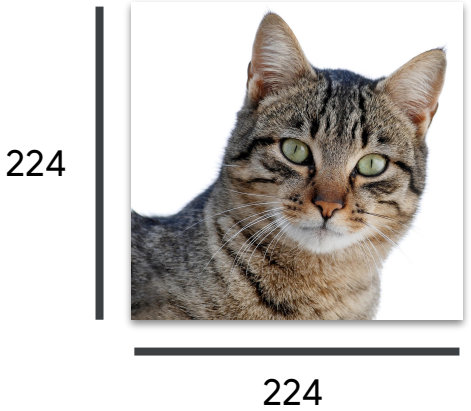
RGB
(# channels)

Bytes/Pixel



Simple **Experiment**

$$224 \times 224 \times 3 \times 4 = 602{,}112 \text{ Bytes}$$

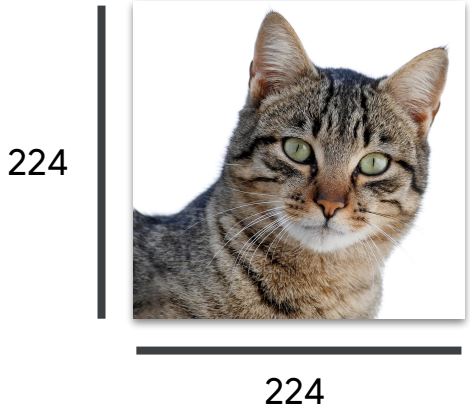


Simple **Experiment**

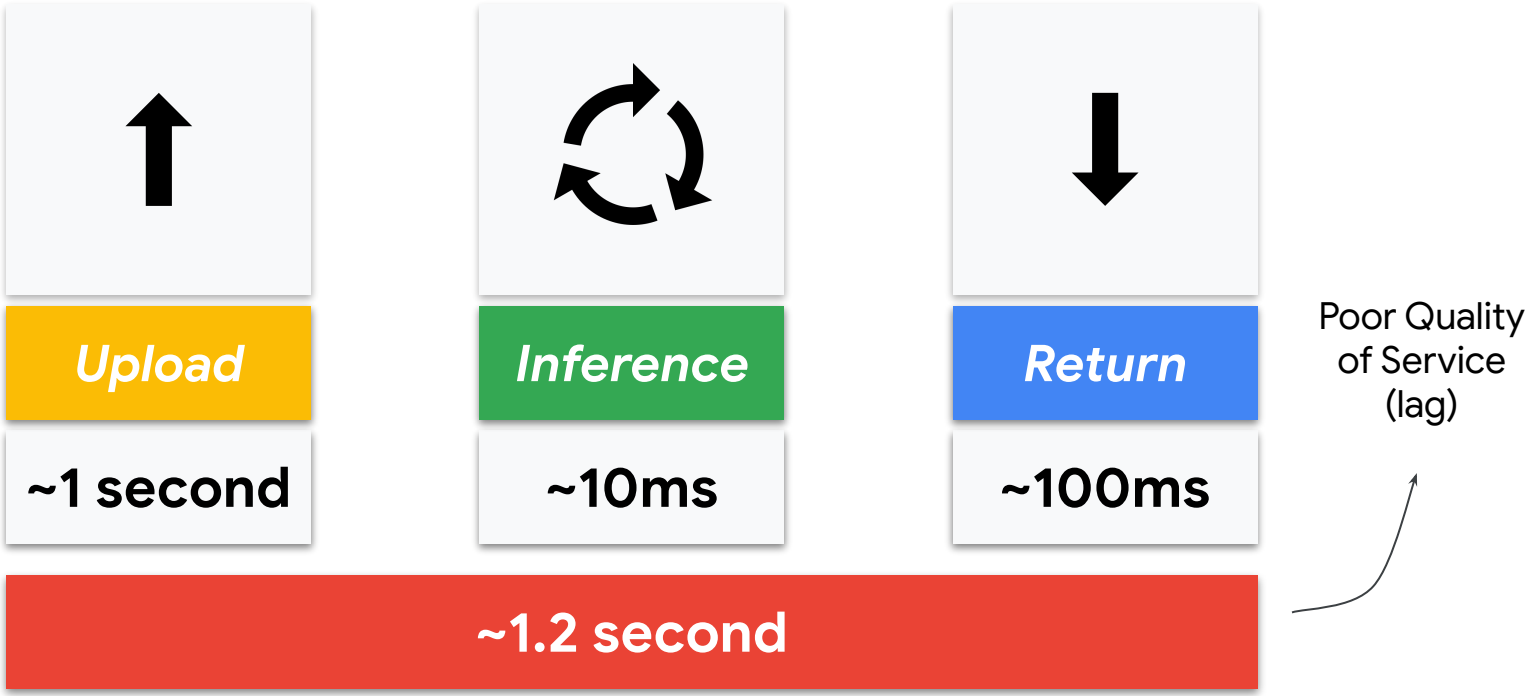
602,112 Bytes

4.6Mbps = 570k ***Bytes*** / Sec

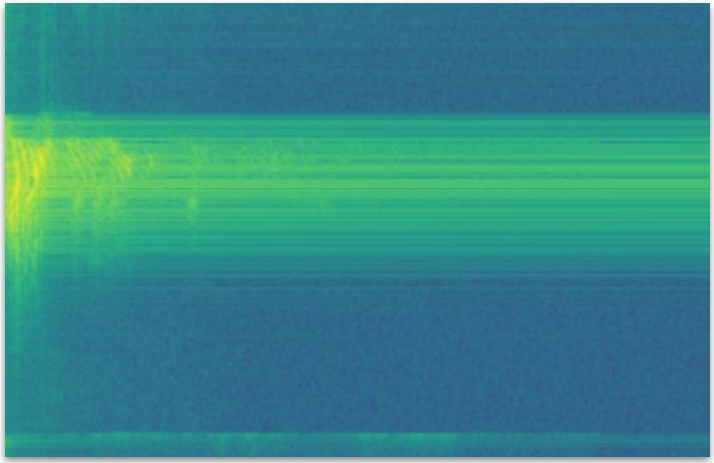
~1 second Transfer Time



Simple **Experiment**



Recall: **Spectrogram**

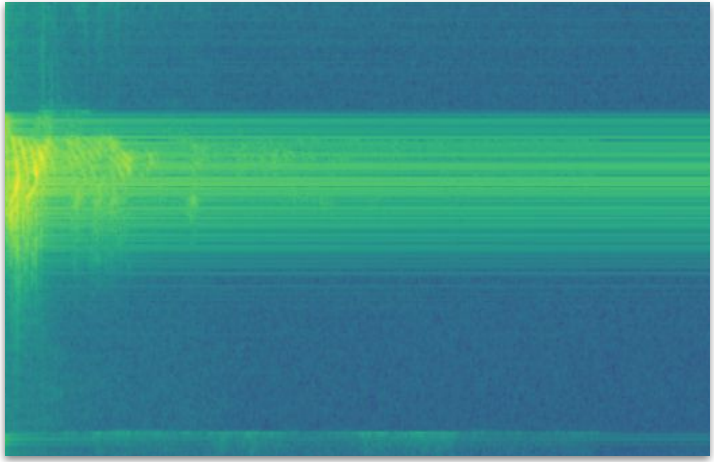


$$49 \times 40 \times 1 \times 4 = 7{,}840 \text{ Bytes}$$

Pixels

RGB
(# channels)

Bytes/Pixel



40

49

$$49 \times 40 \times 1 \times 4 = 7{,}840 \text{ Bytes}$$

Pixels

RGB
(# channels)

Bytes/Pixel

$$224 \times 224 \times 3 \times 4 = 602{,}112 \text{ Bytes}$$

Pixels

RGB
(# channels)

Bytes/Pixel

224

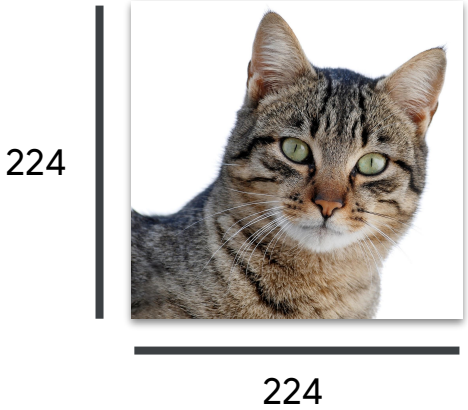
224

Simple **Experiment**

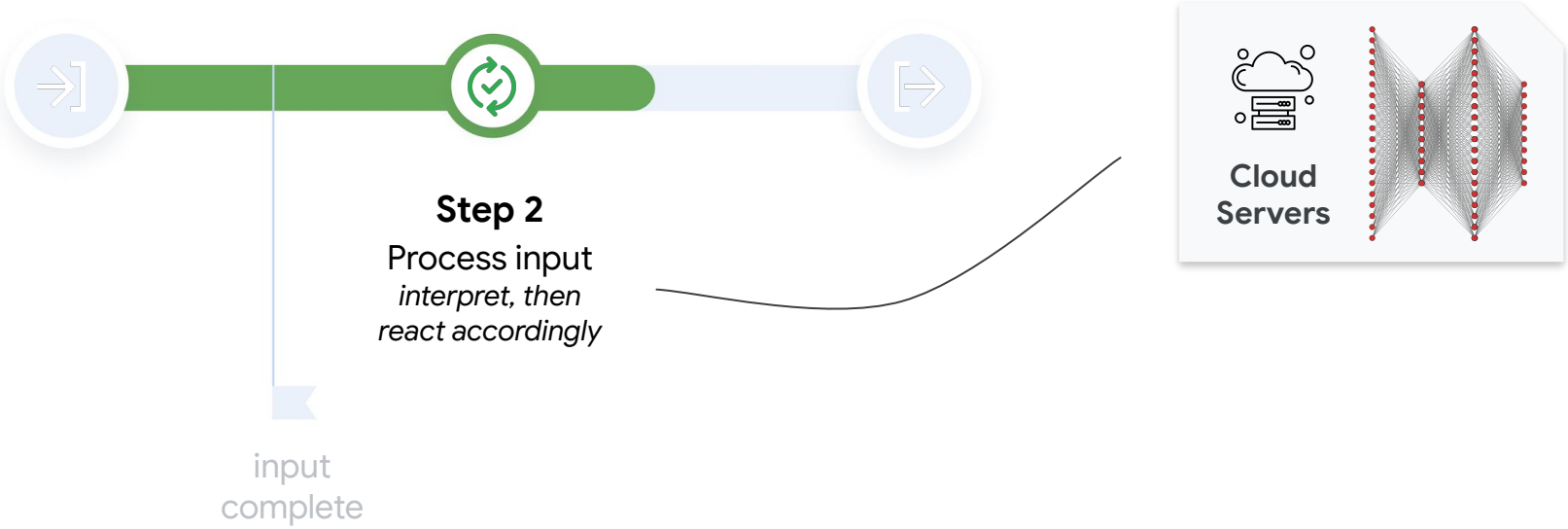
Always-on (Visual Wake Words)?

⟶ Much more data (than KWS)

- Higher **latency**
- Higher **power consumption** (drains battery)
  ⟶ Lower **user satisfaction**



Anatomy of a Visual Wake Words



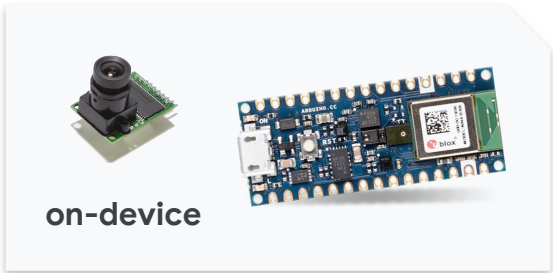
Anatomy of a Visual Wake Words

Step 2

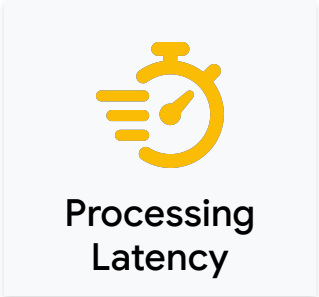
Process input

interpret, then react accordingly

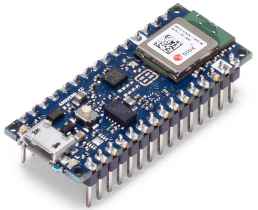
input complete



Constraints for Visual Wake Words



Latency



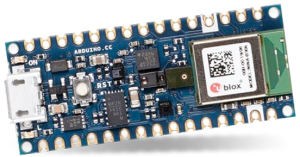
Constraints for Visual Wake Words

Processing Latency

Memory

Memory

Model	Size	Top-1 Accuracy
Xception	**88 MB**	0.790
VGG16	**528 MB**	0.713
ResNet50	**98 MB**	0.749
Inception v3	**92 MB**	0.779
MobileNet v1	**16 MB**	0.713
DenseNet 201	**80 MB**	0.773
NASNetMobile	**23 MB**	0.825

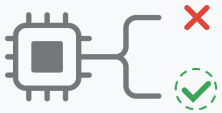


Our board [Course 3 Kit] only has **256 KB** of RAM (memory)

Constraints for Visual Wake Words

Processing Latency

Memory



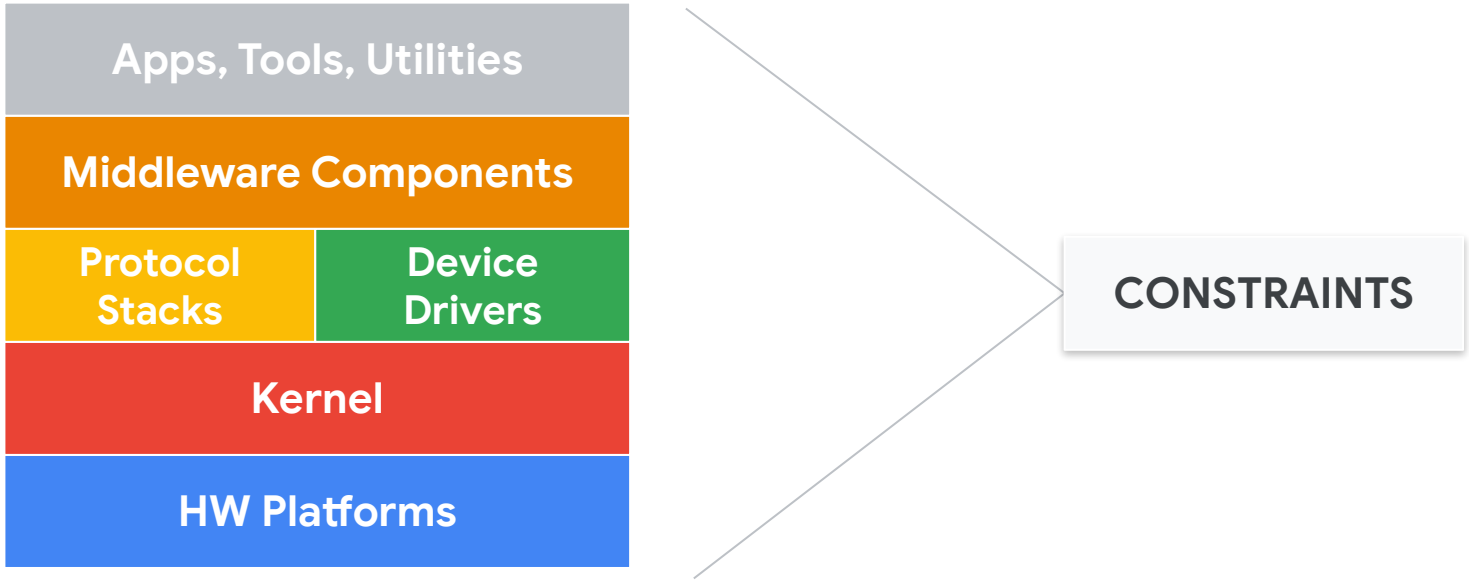
False Positives / Negatives

Errors: False positives/negatives

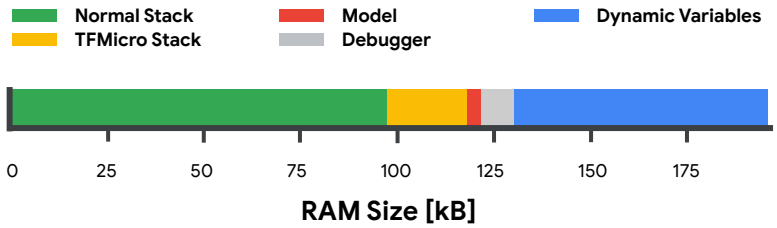
CULTURE | By Moya Lothian-McLean | 29 January 2020, 8:00am

These activists use makeup to defy mass surveillance

Multiple Layers to Compute Stack



Multiple Layers to Compute Stack



Memory: Model + ***Rest of Stack***

Multiple Layers to Compute Stack

Latency: Model + ***Rest of Stack***