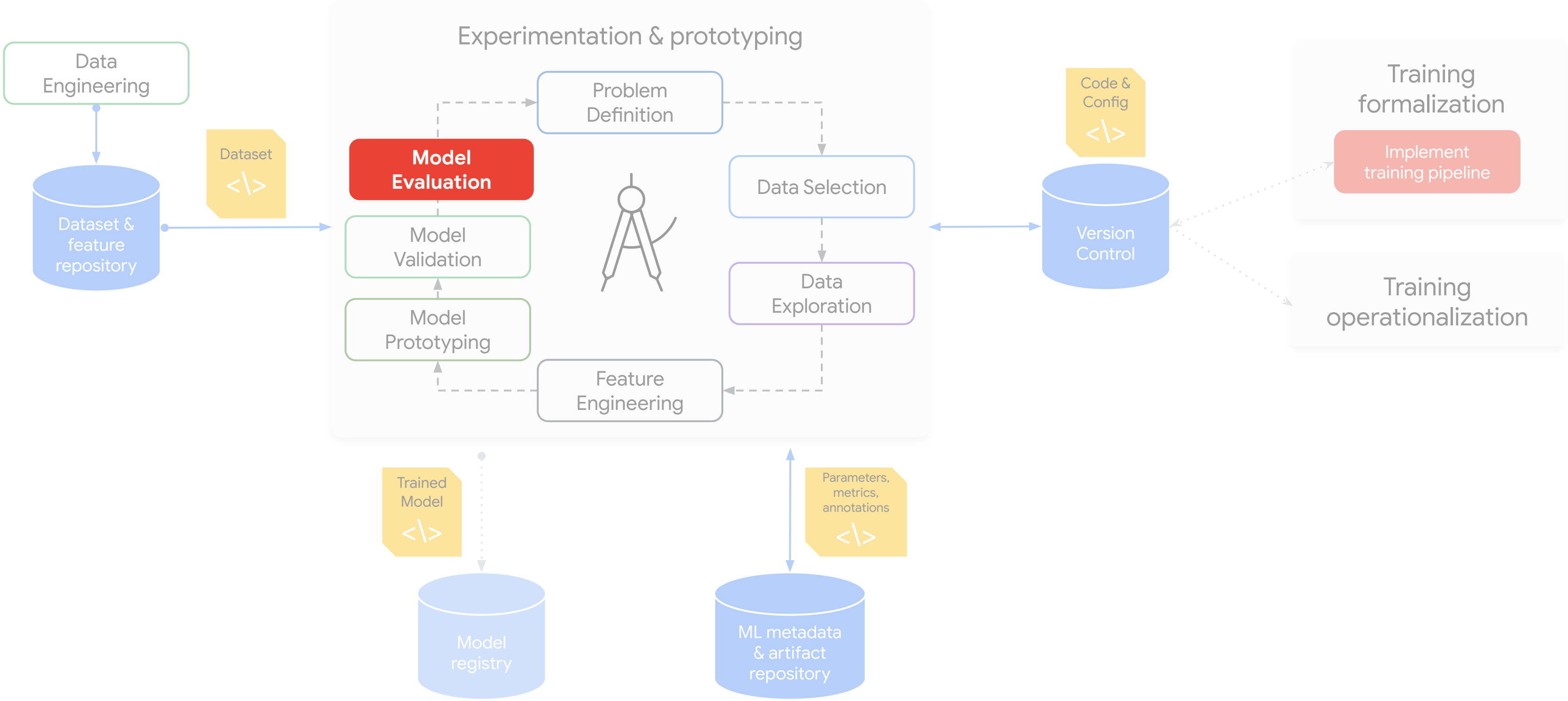
# ML Development: Model Evaluation

# MLOps: ML Development



# The MLOps **Personas**

ML Engineer

**ML Researcher**

Data Scientist

Data Engineer

Software Engineer

DevOps

Business Analyst

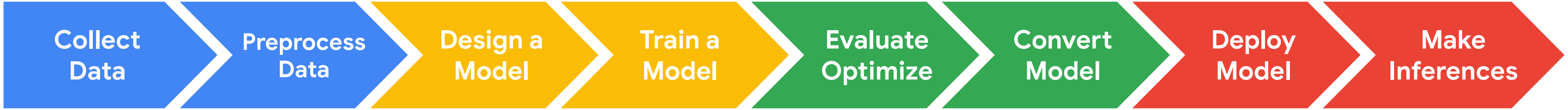
# Constraints for **on-device computing**

Latency

Limited Devices

Battery

# ML Workflow



Collect Data

Preprocess Data

Design a Model

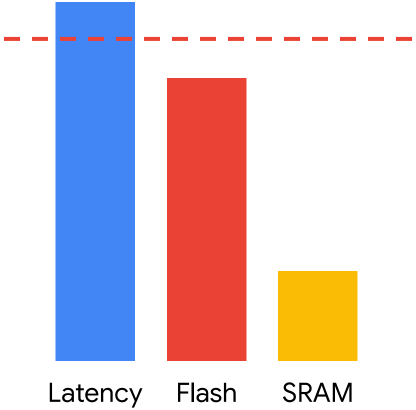
Train a Model

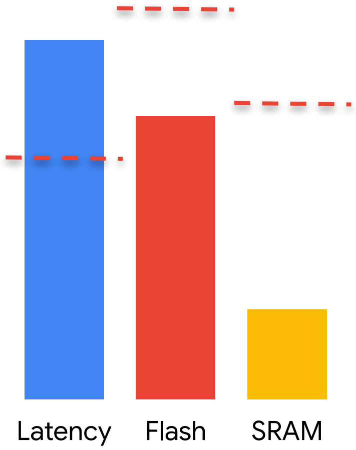
Evaluate Optimize

Convert Model

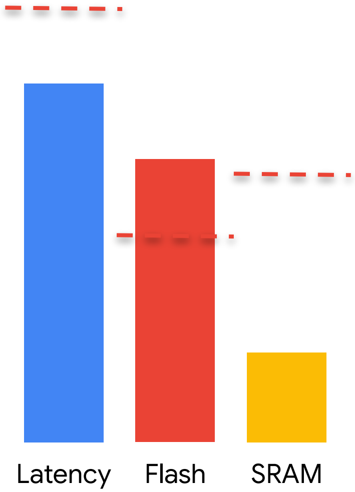
Deploy Model

Make Inferences

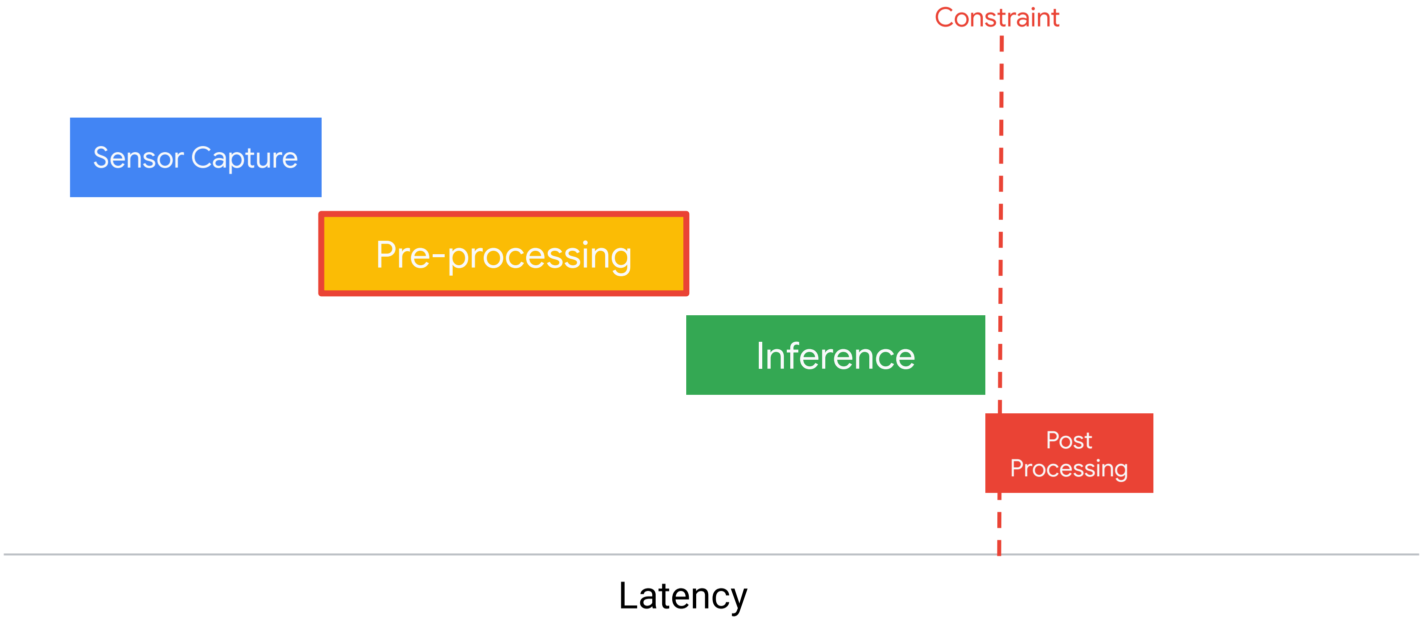


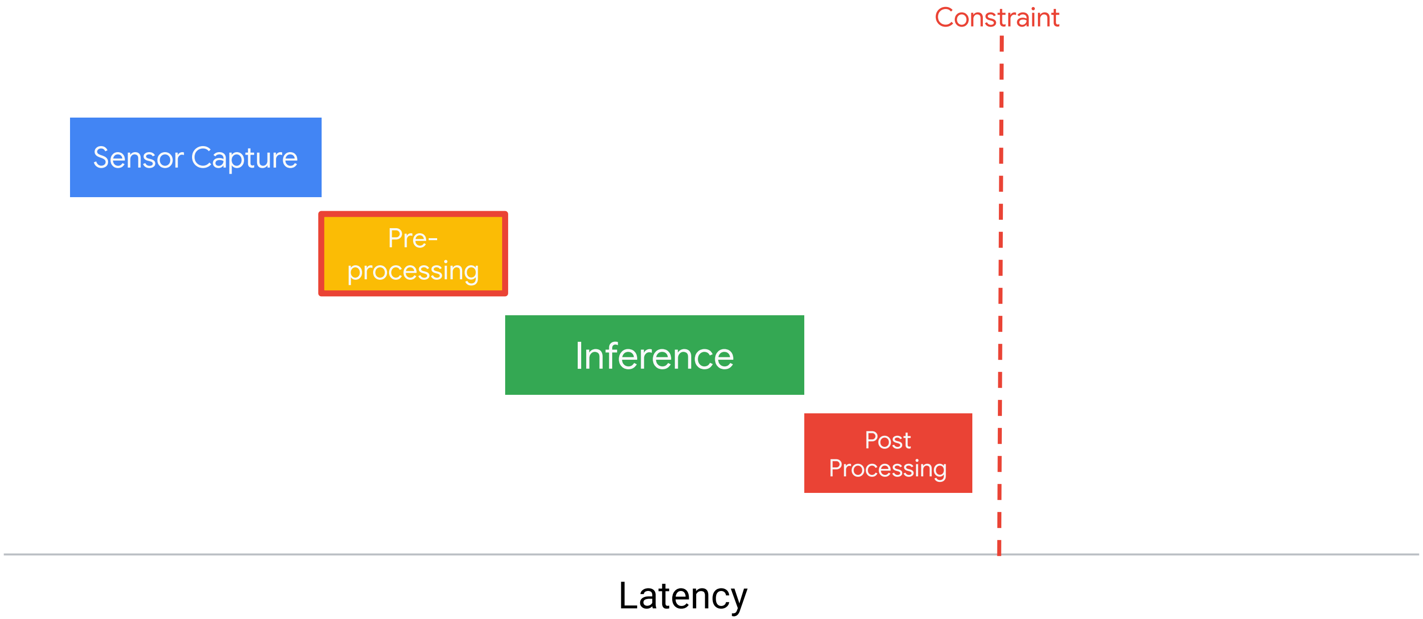


MCU

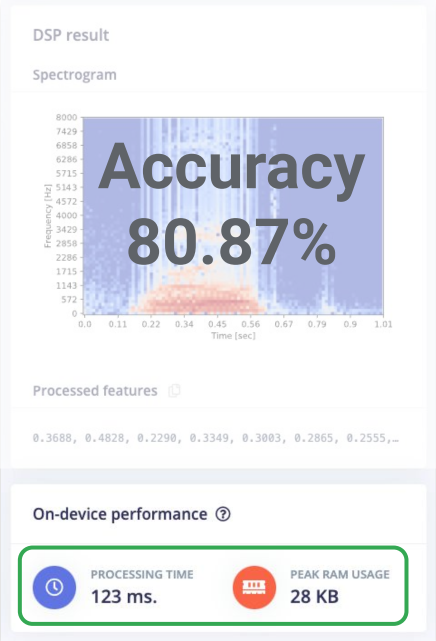


DSP


Constraint
Sensor Capture
Pre-processing
Inference
Post Processing
Latency


Constraint
Sensor Capture
Pre-processing
Inference
Post Processing
Latency

# Spectrograms v. MFCCs



DSP result

Spectrogram

**Accuracy 80.87%**

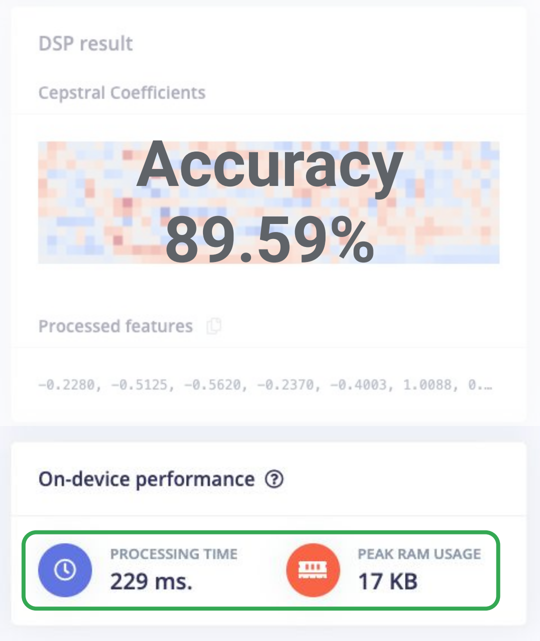
Processed features

```
0.3688, 0.4828, 0.2290, 0.3349, 0.3003, 0.2865, 0.2555,…
```

On-device performance ?

PROCESSING TIME
123 ms.

PEAK RAM USAGE
28 KB



DSP result

Cepstral Coefficients

**Accuracy 89.59%**

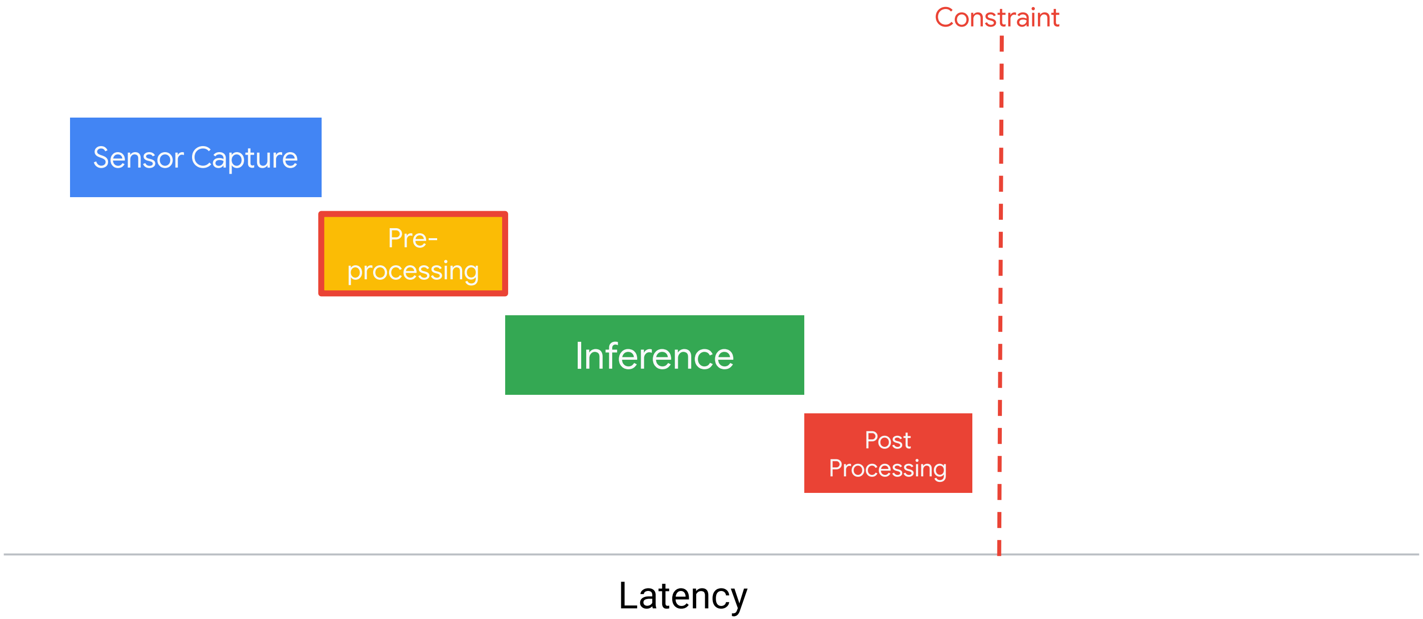
Processed features

```
-0.2280, -0.5125, -0.5620, -0.2370, -0.4003, 1.0088, 0.…
```

On-device performance ?

PROCESSING TIME
229 ms.

PEAK RAM USAGE
17 KB


Constraint
Sensor Capture
Pre-processing
Inference
Post Processing
Latency

# Profiling **Metrics**

Latency

Memory

Energy

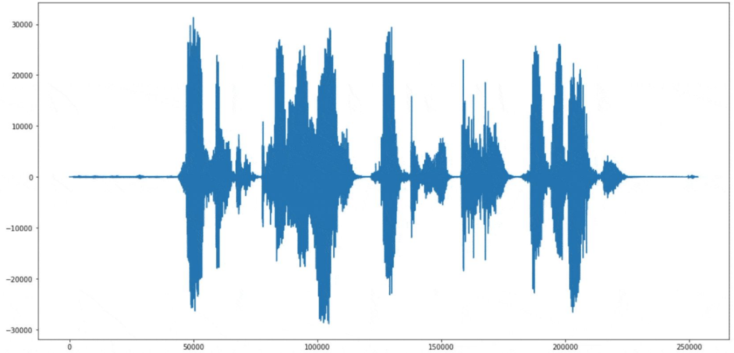
# Profiling **Metrics**

**Latency**

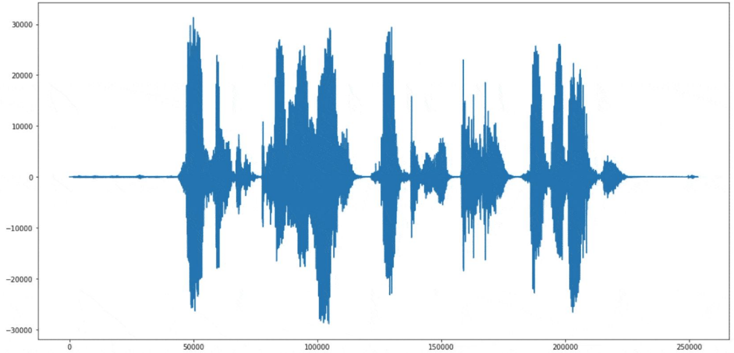
Memory

Energy

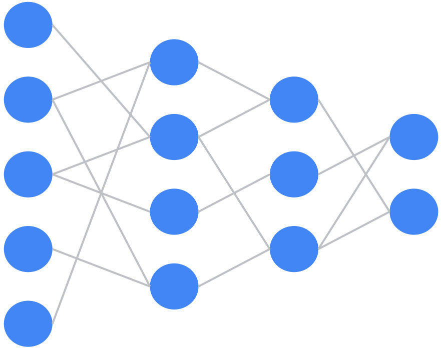
**Desired**
Latency



**Deployed**
Latency

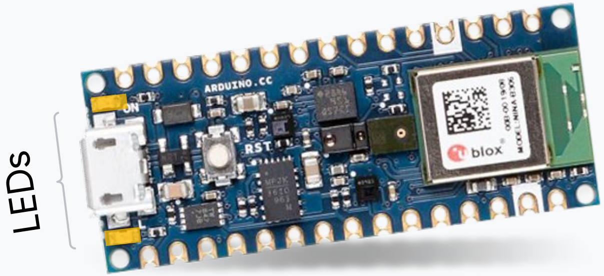


# Operation Count



Math OPs ~ Latency

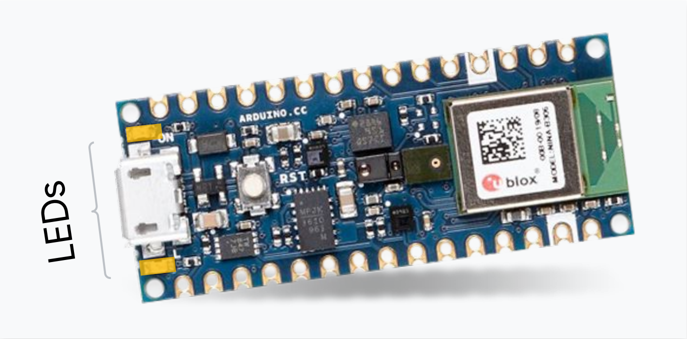
# Latency Profiling Methods



Internal Timer

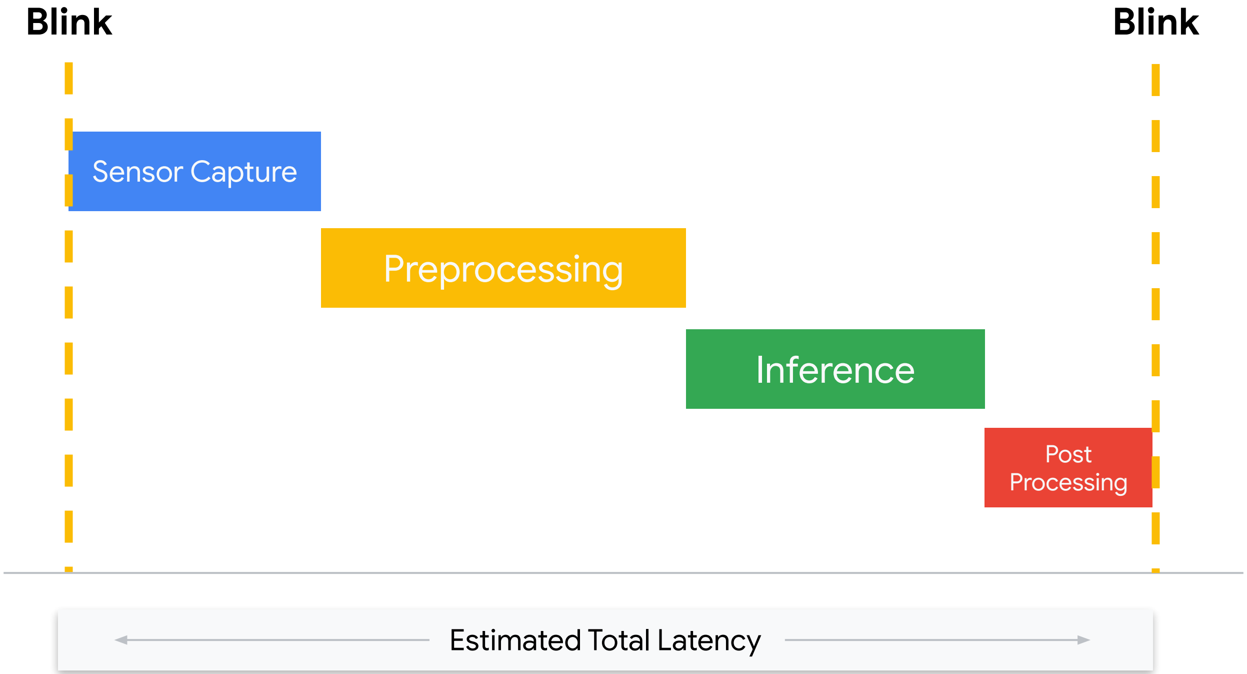
other tools

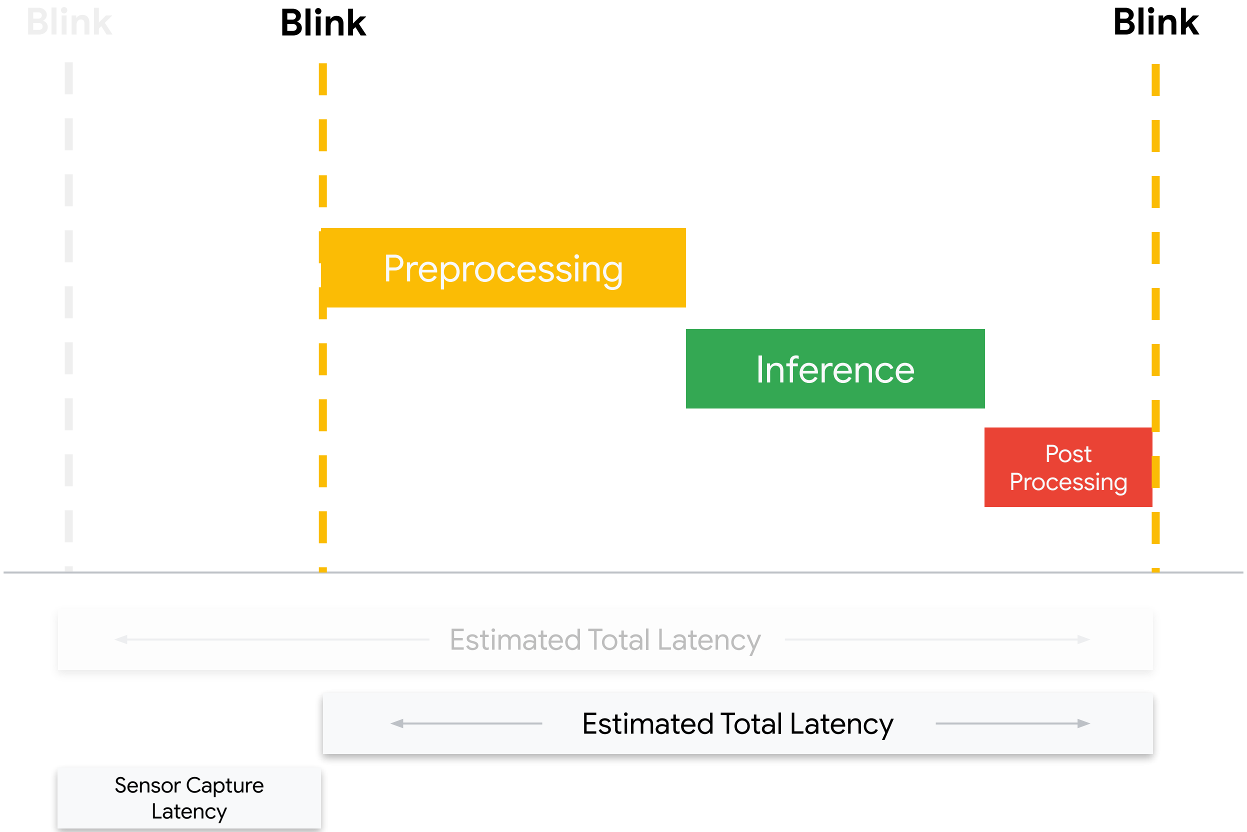
# Latency Profiling Methods



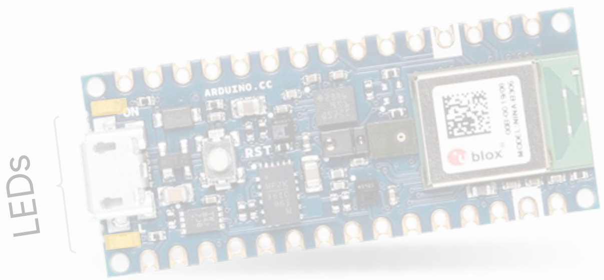
Internal Timer

other tools


Blink
Blink
Sensor Capture
Preprocessing
Inference
Post Processing
Estimated Total Latency

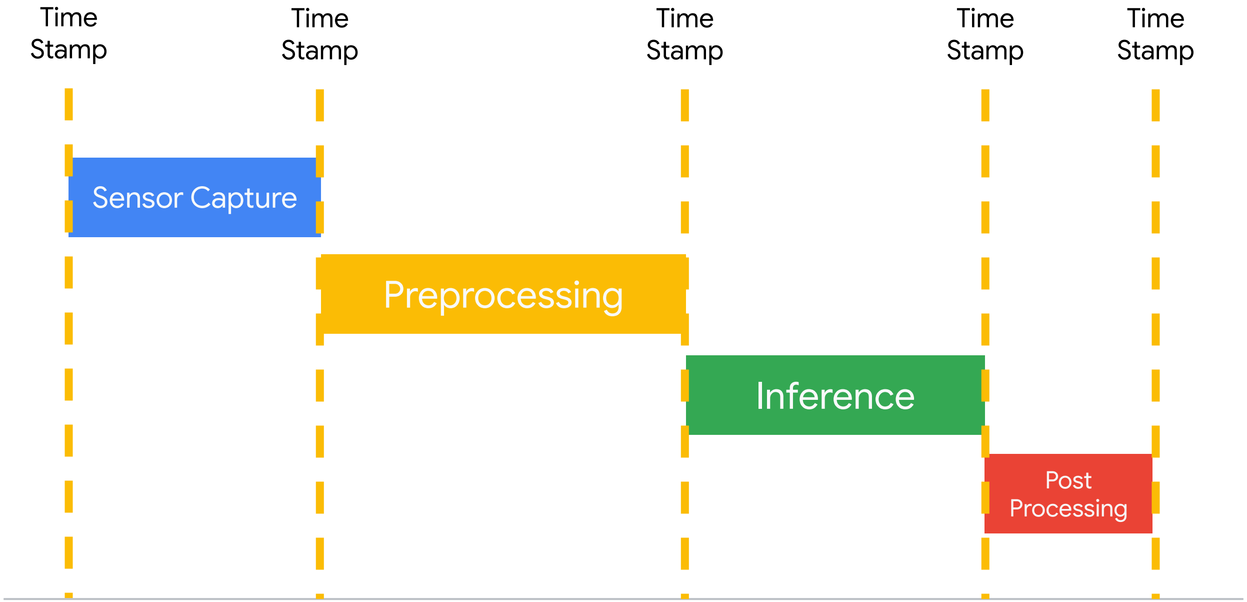

Blink
Blink
Blink
Preprocessing
Inference
Post Processing
Estimated Total Latency
Estimated Total Latency
Sensor Capture Latency

# Latency Profiling Methods

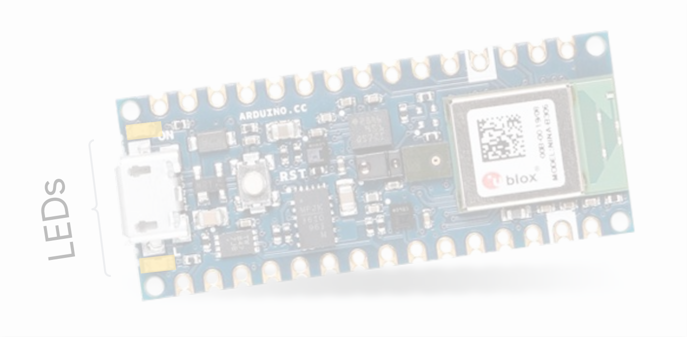


Internal Timer

other tools


Time Stamp
Time Stamp
Time Stamp
Time Stamp
Time Stamp
Sensor Capture
Preprocessing
Inference
Post Processing

# Latency Profiling Methods

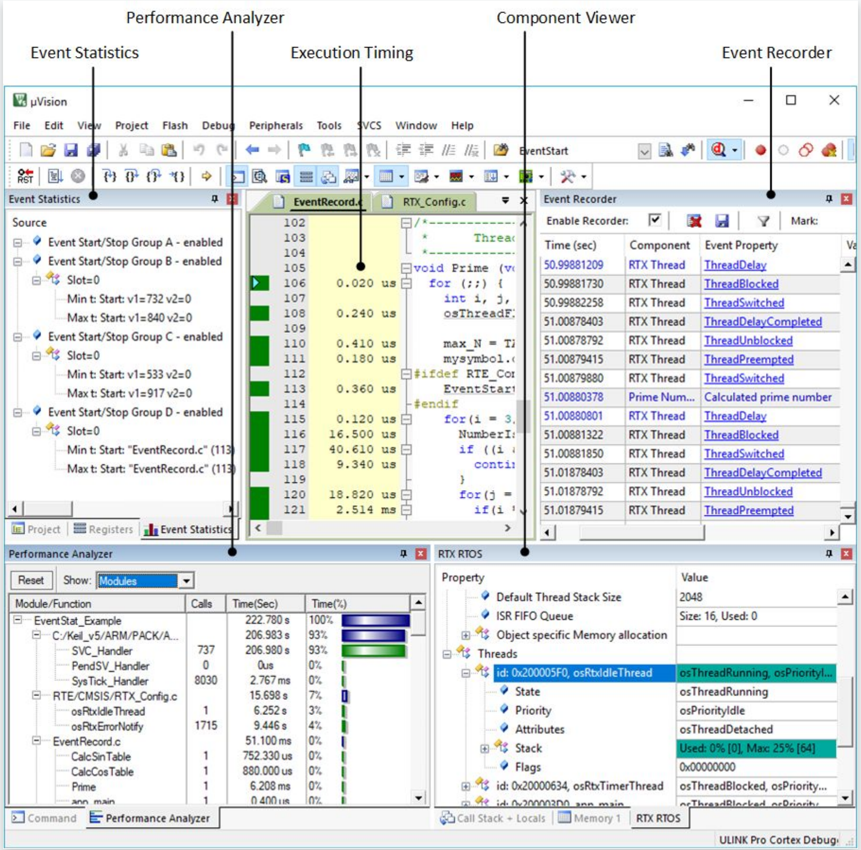


Internal Timer

other tools

# Profiling **Tools**

- More advanced tools can be used to understand the latency impact at a very **Fine grained level**.



Event Statistics

```
Source
⊟ Event Start/Stop Group A - enabled
⊟ Event Start/Stop Group B - enabled
    ⊟ Slot=0
        Min t: Start: v1=732 v2=0
        Max t: Start: v1=840 v2=0
⊟ Event Start/Stop Group C - enabled
    ⊟ Slot=0
        Min t: Start: v1=533 v2=0
        Max t: Start: v1=917 v2=0
⊟ Event Start/Stop Group D - enabled
    ⊟ Slot=0
        Min t: Start: "EventRecord.c" (113
        Max t: Start: "EventRecord.c" (113
```

```
             /*------------
              *        Threa
              *------------
             void Prime (vo
 0.020 us      for (;;) {
                 int i, j,
 0.240 us        osThreadF
 
 0.410 us        max_N = TA
 0.180 us        mysymbol.
             #ifdef RTE_Co
 0.360 us        EventStart
             #endif
 0.120 us        for(i = 3,
16.500 us          NumberI
40.610 us          if ((i
 9.340 us            conti
                   }
18.820 us          for(j =
 2.514 ms            if(i
```

Event Recorder

Enable Recorder: ☑ Mark:

| Time (sec) | Component | Event Property | Va |
|---|---|---|---|
| 50.99881209 | RTX Thread | ThreadDelay | |
| 50.99881730 | RTX Thread | ThreadBlocked | |
| 50.99882258 | RTX Thread | ThreadSwitched | |
| 51.00878403 | RTX Thread | ThreadDelayCompleted | |
| 51.00878792 | RTX Thread | ThreadUnblocked | |
| 51.00879415 | RTX Thread | ThreadPreempted | |
| 51.00879880 | RTX Thread | ThreadSwitched | |
| 51.00880378 | Prime Num... | Calculated prime number | |
| 51.00880801 | RTX Thread | ThreadDelay | |
| 51.00881322 | RTX Thread | ThreadBlocked | |
| 51.00881850 | RTX Thread | ThreadSwitched | |
| 51.01878403 | RTX Thread | ThreadDelayCompleted | |
| 51.01878792 | RTX Thread | ThreadUnblocked | |
| 51.01879415 | RTX Thread | ThreadPreempted | |

Performance Analyzer

| Module/Function | Calls | Time(Sec) | Time(%) |
|---|---|---|---|
| ⊟ EventStat_Example | | 222.780 s | 100% |
| ⊟ C:/Keil_v5/ARM/PACK/A... | | 206.983 s | 93% |
| SVC_Handler | 737 | 206.980 s | 93% |
| PendSV_Handler | 0 | 0us | 0% |
| SysTick_Handler | 8030 | 2.767 ms | 0% |
| ⊟ RTE/CMSIS/RTX_Config.c | | 15.698 s | 7% |
| osRtxIdleThread | 1 | 6.252 s | 3% |
| osRtxErrorNotify | 1715 | 9.446 s | 4% |
| ⊟ EventRecord.c | | 51.100 ms | 0% |
| CalcSinTable | 1 | 752.330 us | 0% |
| CalcCosTable | 1 | 880.000 us | 0% |
| Prime | 1 | 6.208 ms | 0% |
| app_main | 1 | 0.400 us | 0% |

RTX RTOS

| Property | Value |
|---|---|
| Default Thread Stack Size | 2048 |
| ISR FIFO Queue | Size: 16, Used: 0 |
| ⊞ Object specific Memory allocation | |
| ⊟ Threads | |
| ⊟ id: 0x200005F0, osRtxIdleThread | osThreadRunning, osPriorityI... |
| State | osThreadRunning |
| Priority | osPriorityIdle |
| Attributes | osThreadDetached |
| ⊞ Stack | Used: 0% [0], Max: 25% [64] |
| Flags | 0x00000000 |
| ⊞ id: 0x20000634, osRtxTimerThread | osThreadBlocked, osPriority... |
| ⊞ id: 0x200003D0, app_main | osThreadBlocked, osPriority... |

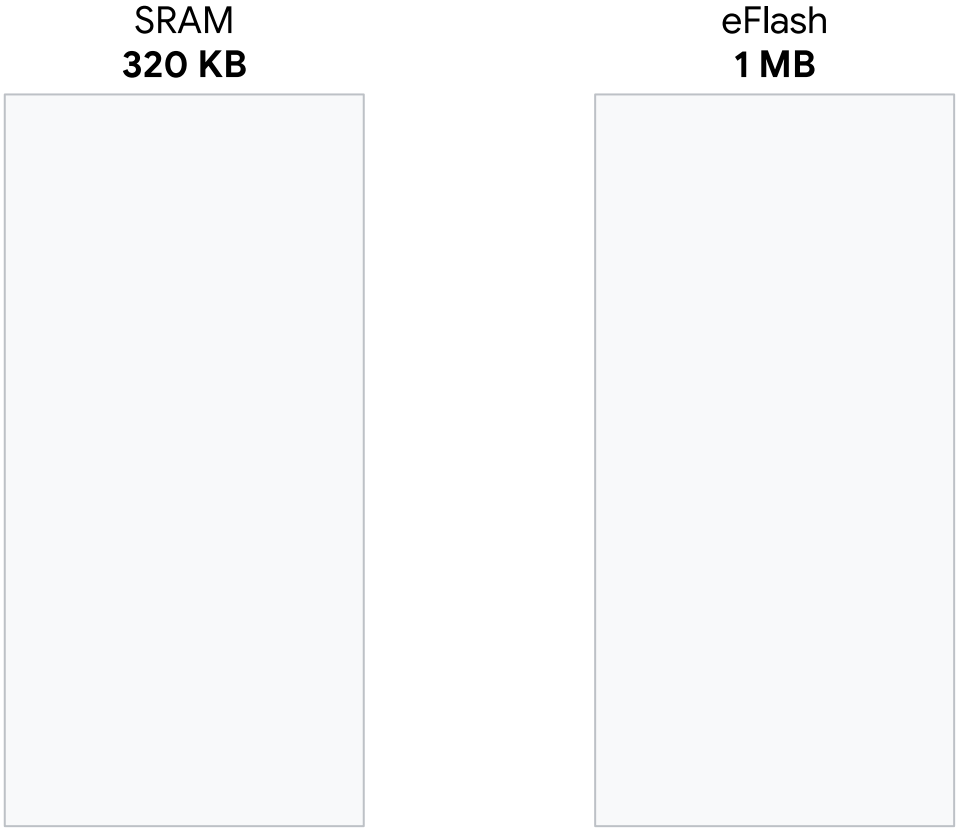
# Profiling **Metrics**

Latency

**Memory**

Energy

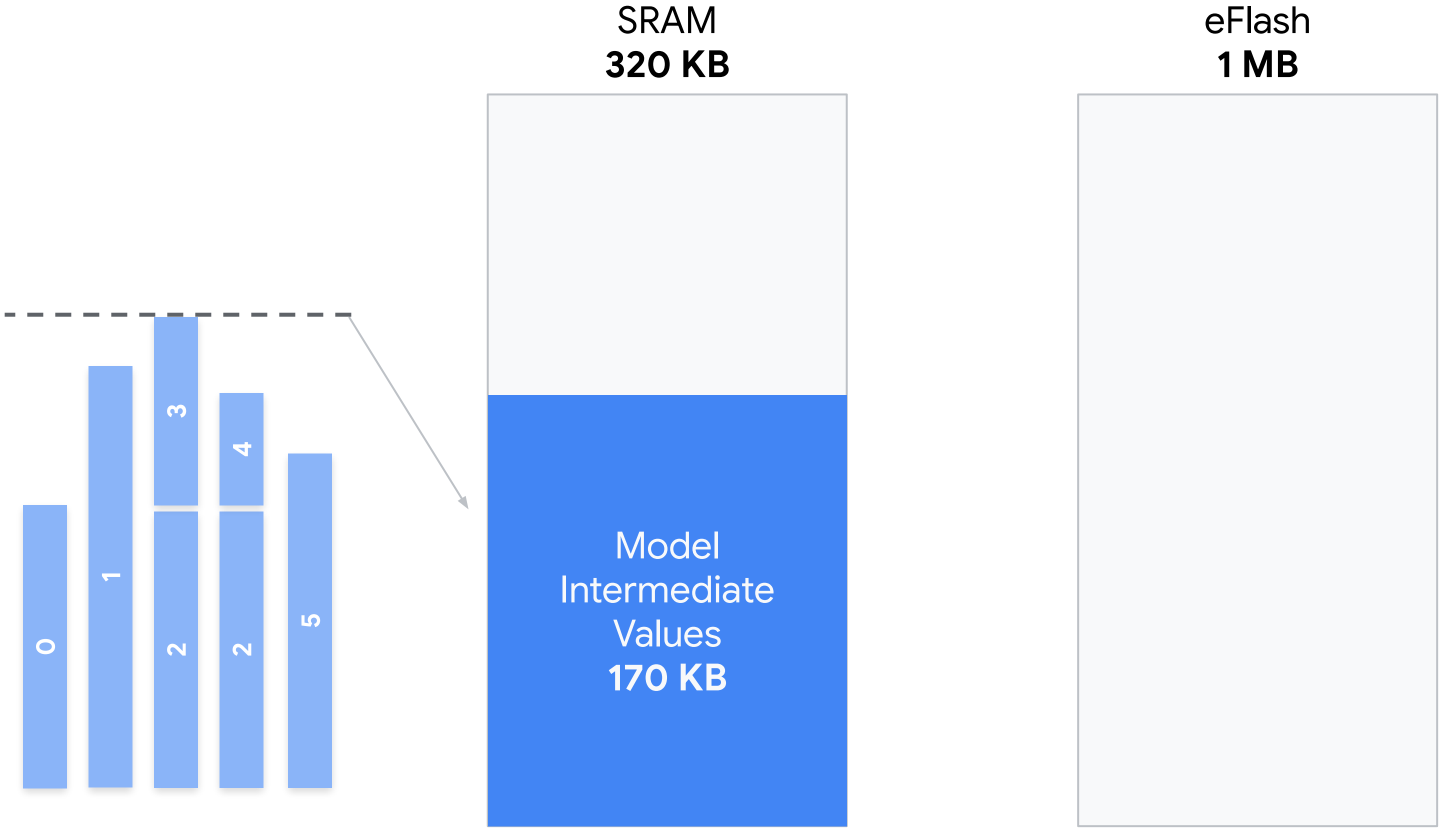
# Memory and Storage



SRAM
**320 KB**

eFlash
**1 MB**

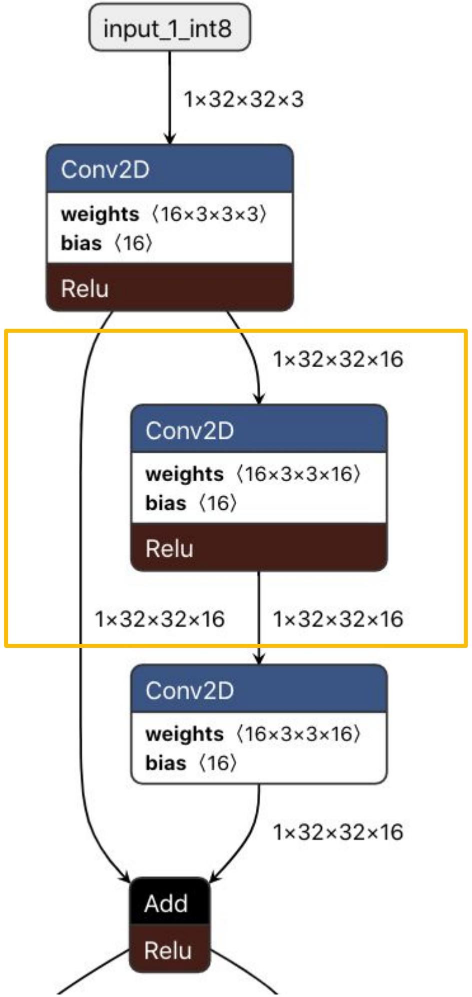
# Memory and Storage



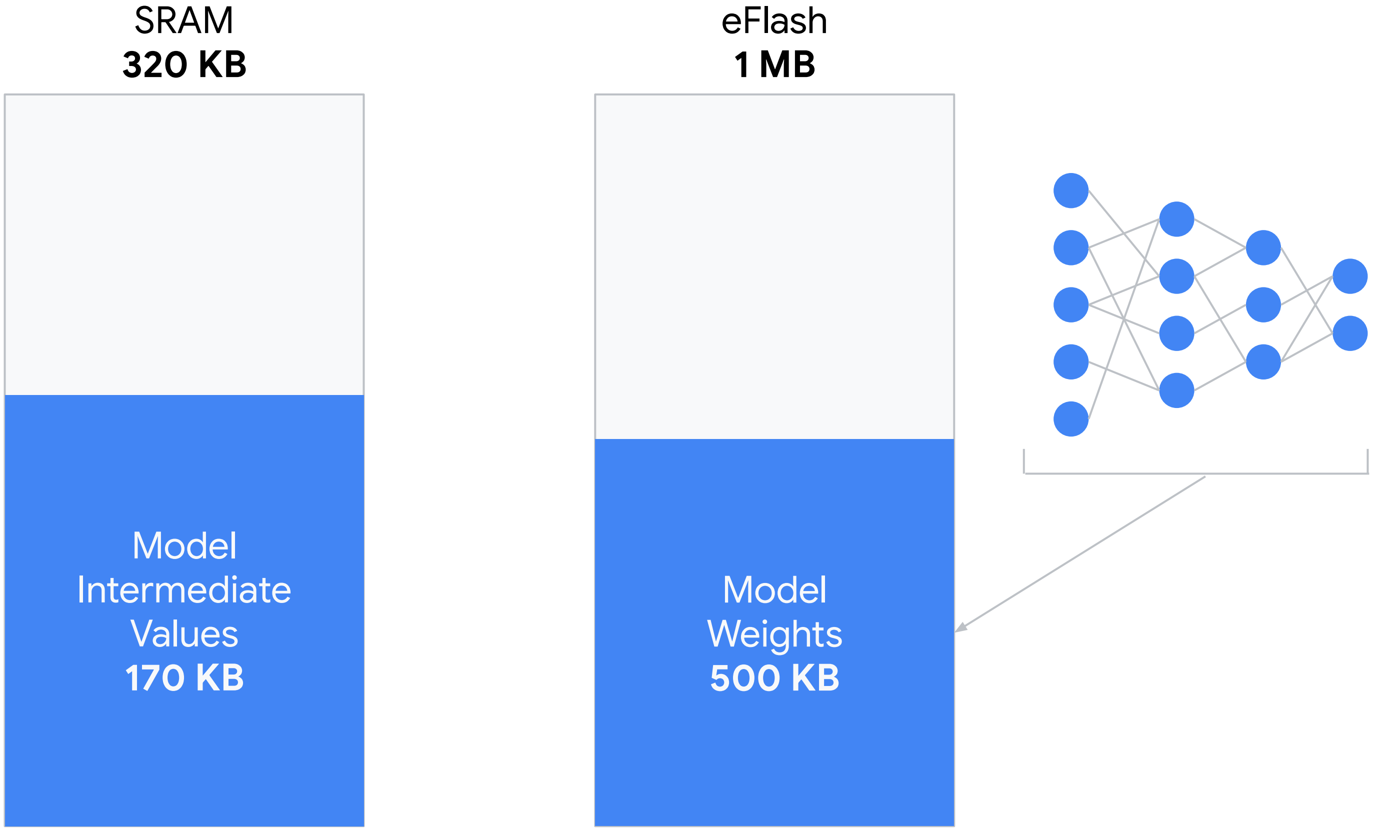
SRAM
**320 KB**

Model
Intermediate
Values
**170 KB**

eFlash
**1 MB**


input_1_int8
1×32×32×3
Conv2D
weights ⟨16×3×3×3⟩
bias ⟨16⟩
Relu
1×32×32×16
Conv2D
weights ⟨16×3×3×16⟩
bias ⟨16⟩
Relu
1×32×32×16
1×32×32×16
Conv2D
weights ⟨16×3×3×16⟩
bias ⟨16⟩
1×32×32×16
Add
Relu

# Memory and Storage



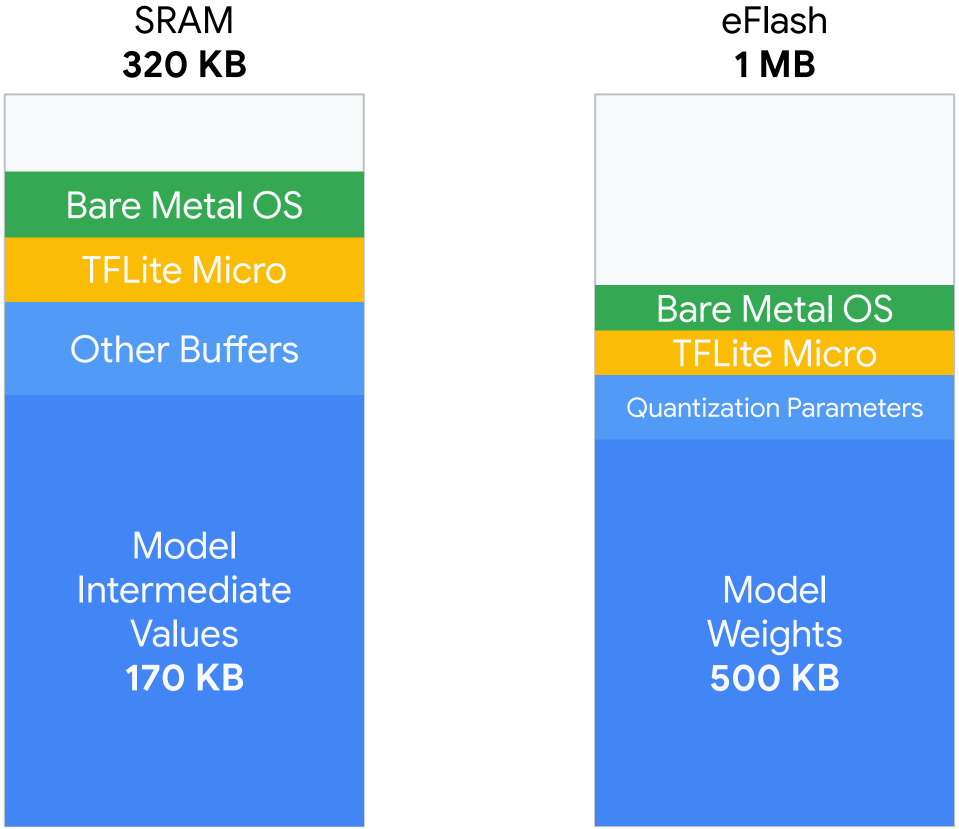
SRAM
**320 KB**

Model
Intermediate
Values
**170 KB**

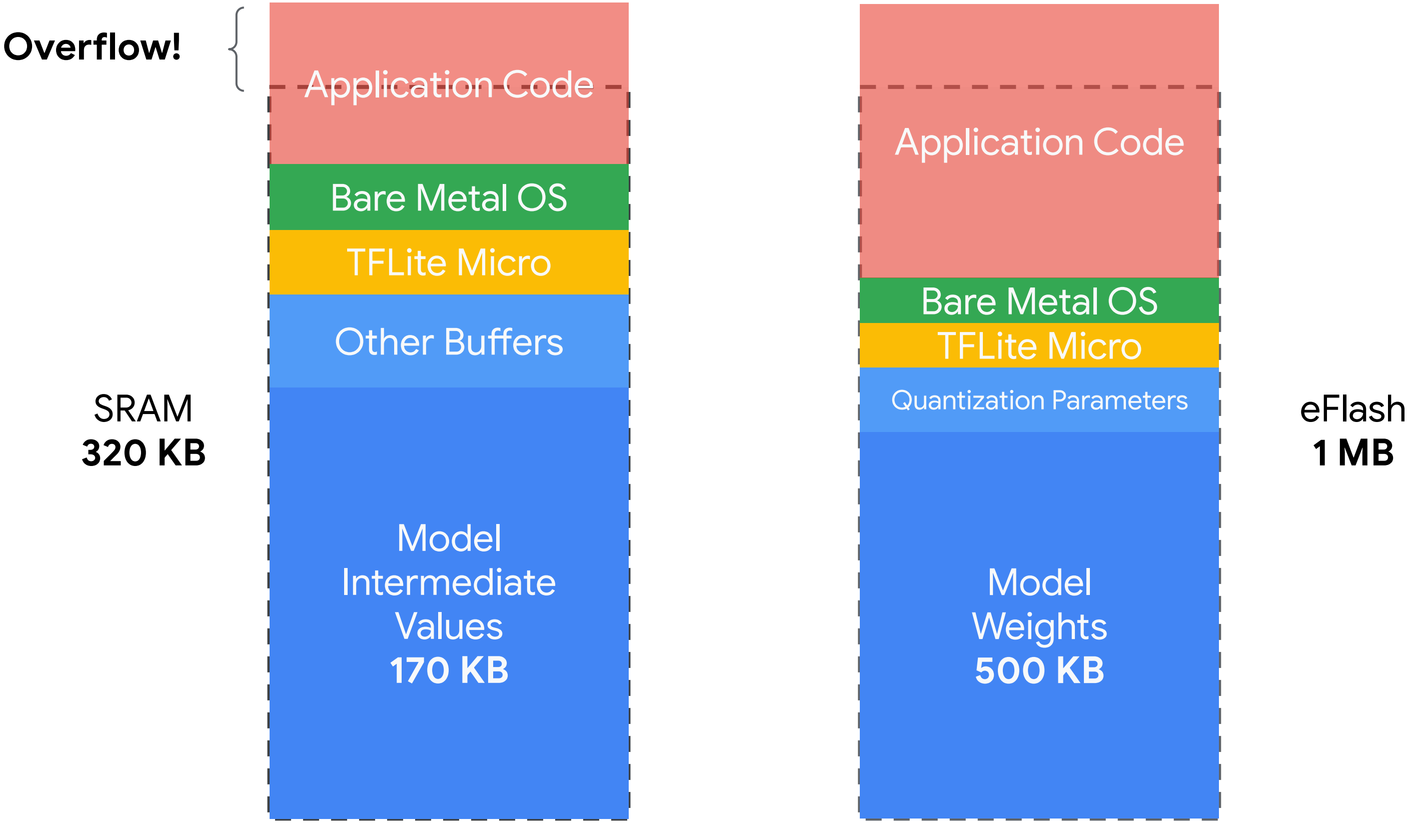
eFlash
**1 MB**

Model
Weights
**500 KB**

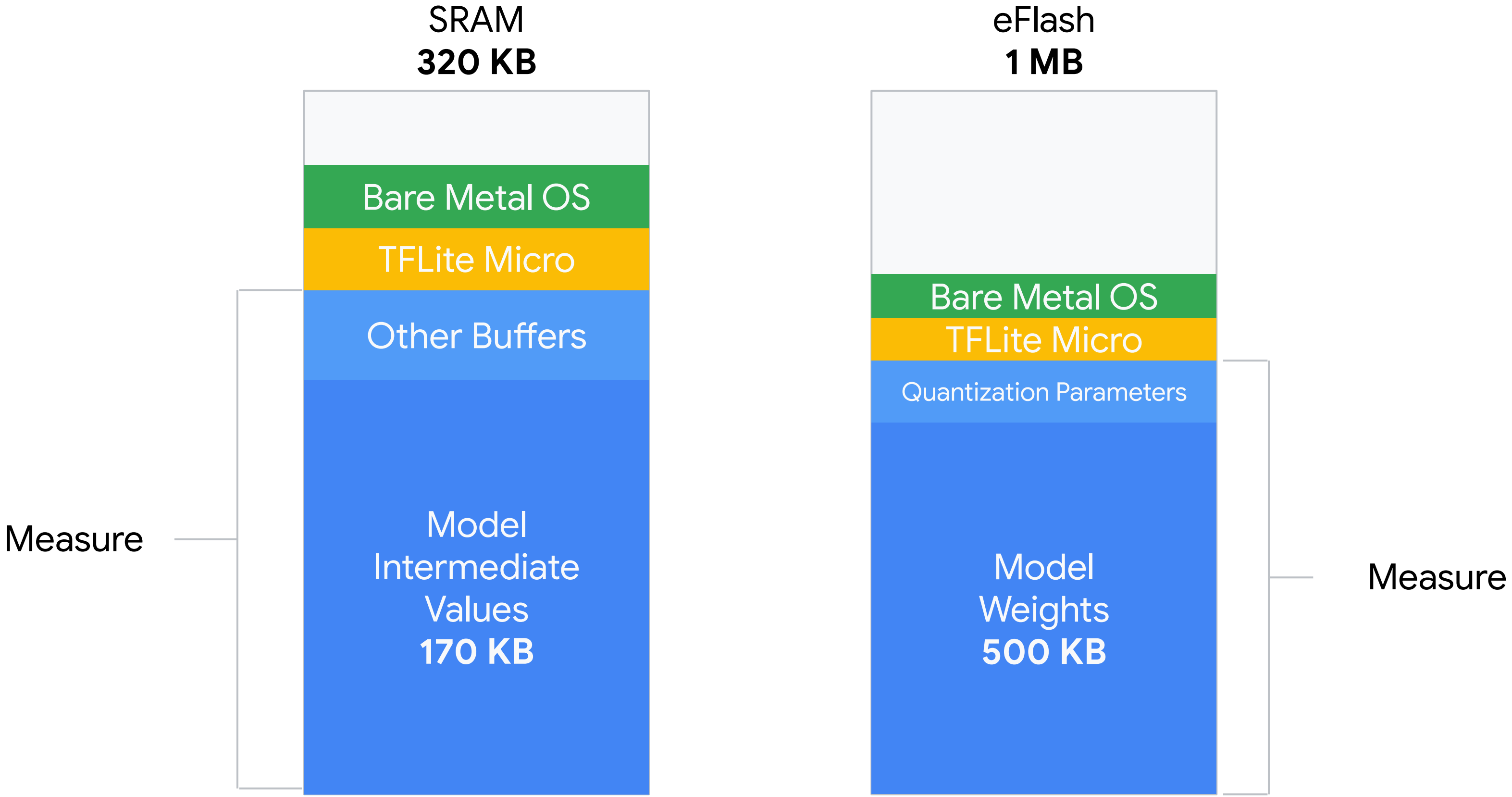
# Memory and Storage



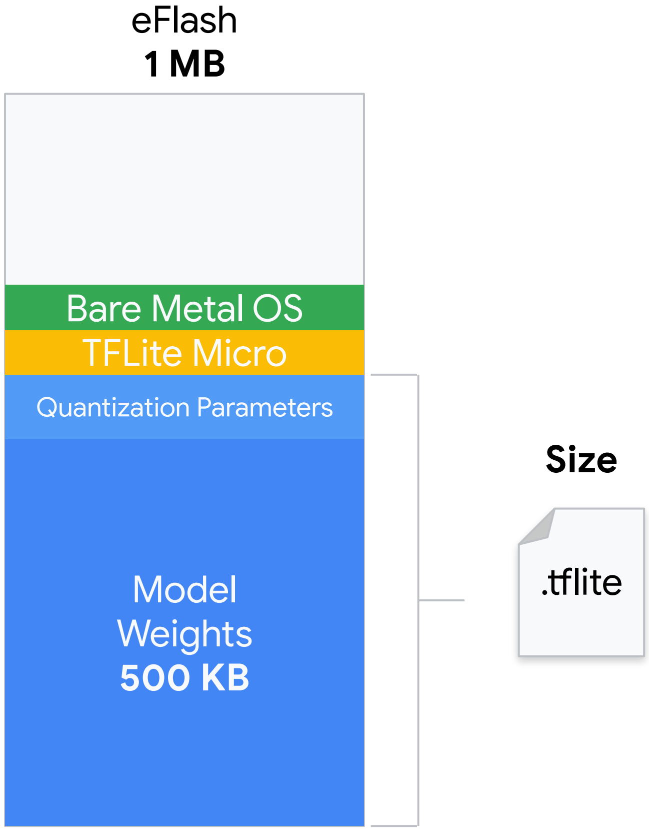
# Memory and Storage



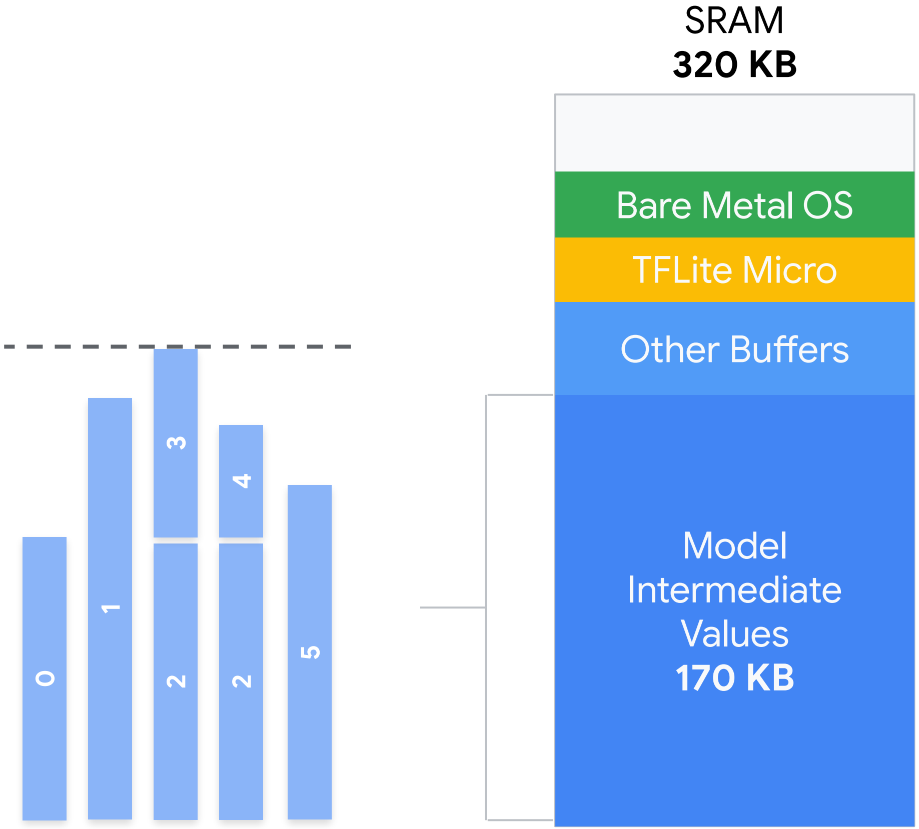
# Memory and Storage



# Memory and Storage



# Memory and Storage



SRAM
**320 KB**

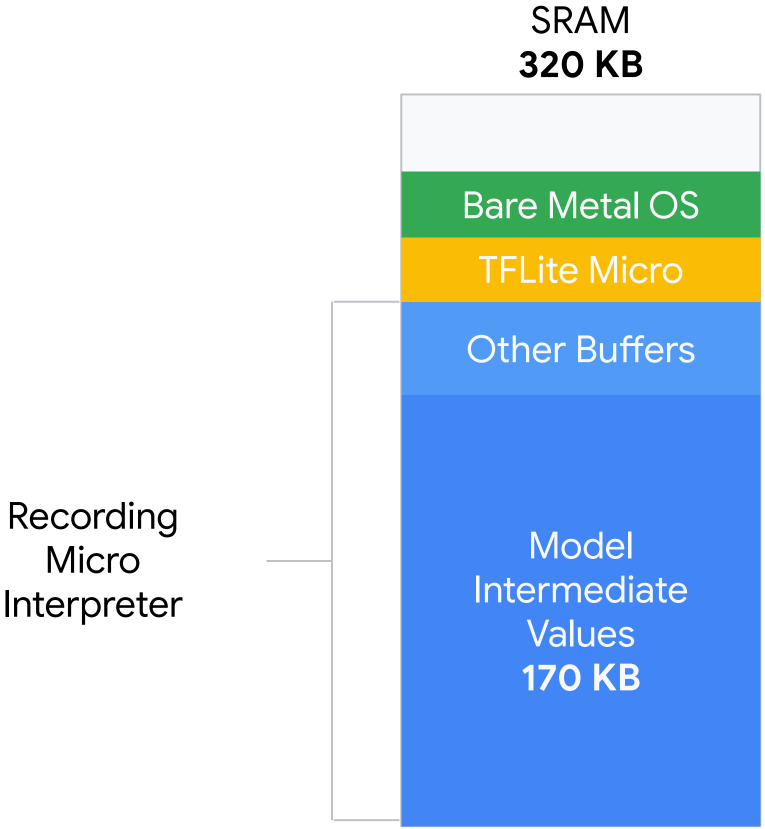
Bare Metal OS

TFLite Micro

Other Buffers

Model
Intermediate
Values
**170 KB**

# Memory and Storage



# Profiling Metrics

Latency

Memory

Energy

# **Estimating** Energy

Estimate by **latency**

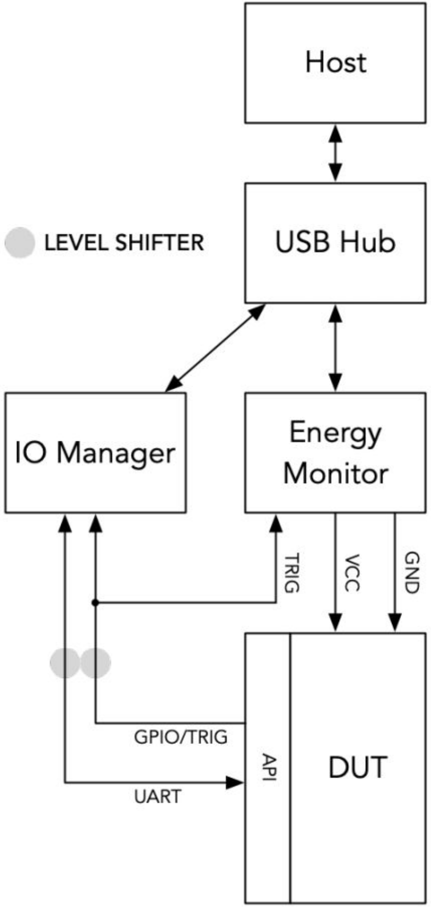
- Rough estimate
- Only relative to other models

Latency

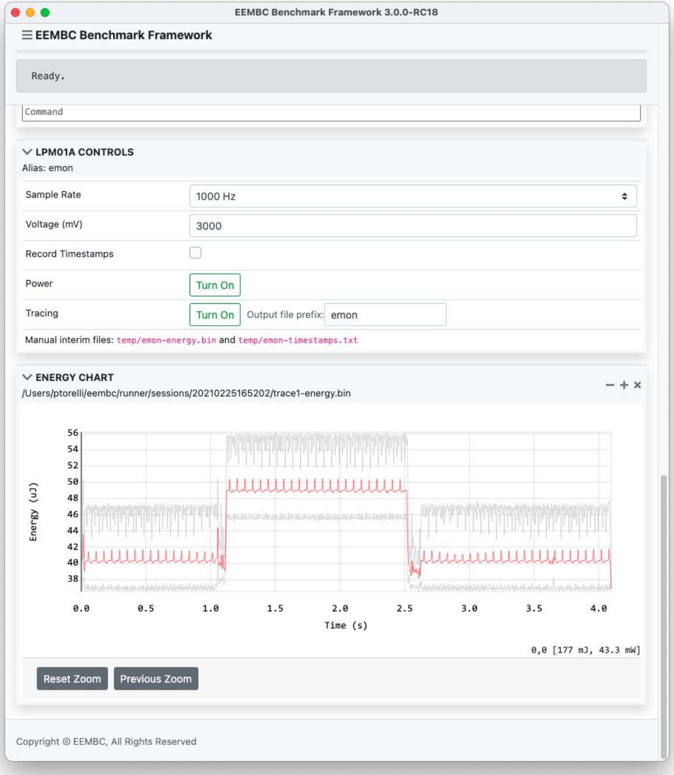
# **Measuring** Energy

Accurately measuring energy is **complex**:

- Isolating out I/O and power planes can be complicated



# EEMBC's EnergyRunner™



EEMBC Benchmark Framework 3.0.0-RC18

☰ EEMBC Benchmark Framework

```
Ready.
```

Command

LPM01A CONTROLS

Alias: emon

| | |
|---|---|
| Sample Rate | 1000 Hz |
| Voltage (mV) | 3000 |
| Record Timestamps | □ |
| Power | Turn On |
| Tracing | Turn On  Output file prefix: emon |

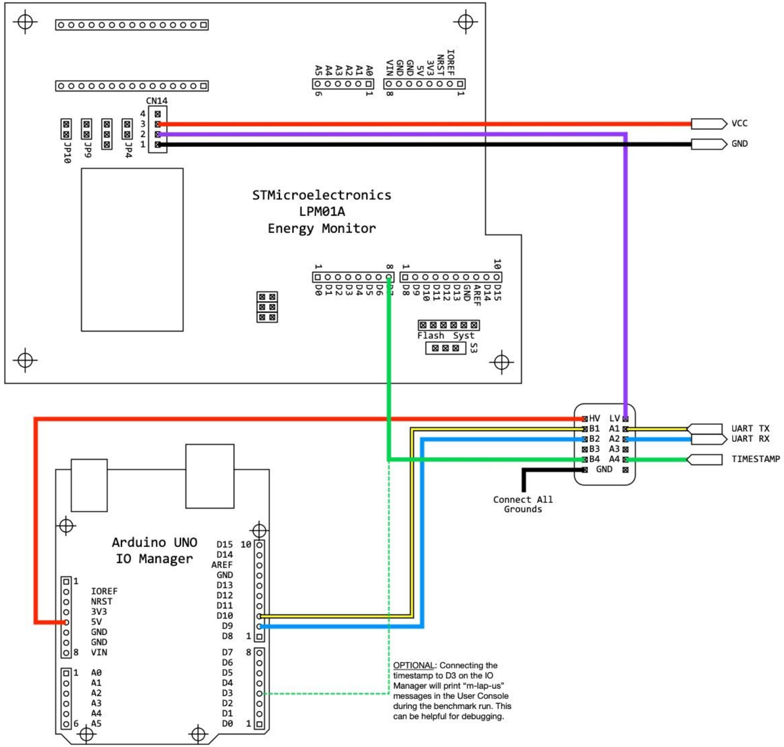
Manual interim files: `temp/emon-energy.bin` and `temp/emon-timestamps.txt`

ENERGY CHART

/Users/ptorelli/eembc/runner/sessions/20210225165202/trace1-energy.bin

```
0,0 [177 mJ, 43.3 mW]
```

Copyright © EEMBC, All Rights Reserved

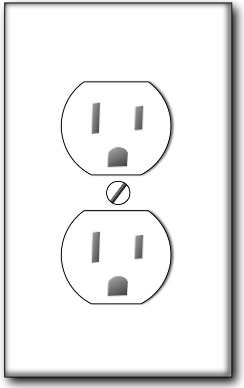


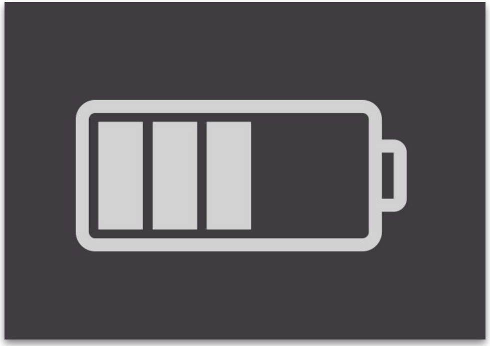
OPTIONAL: Connecting the timestamp to D3 on the IO Manager will print "m-lap-us" messages in the User Console during the benchmark run. This can be helpful for debugging.

# **Deployment** Scenario

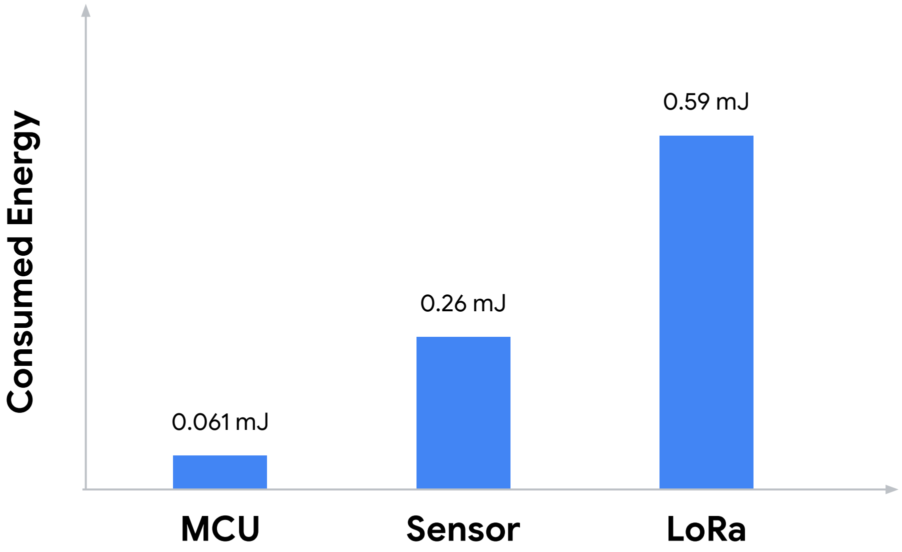
**Battery** Powered:

- Size of battery?
- How often is it charged?
- Energy Harvesting?





# Other Factors



# Constraints for **on-device computing**

Latency

Limited Devices

Battery