2021 embedded vision summit

VIRTUAL | MAY 25-28

Hate it or love it, your SW stack defines application performance and reach

Felix Baum, Director, Product Management at Qualcomm Technologies Inc.

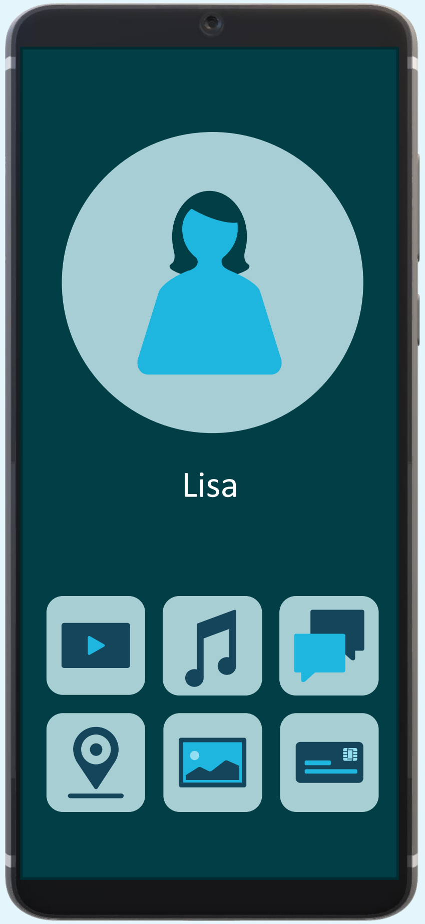
Qualcomm

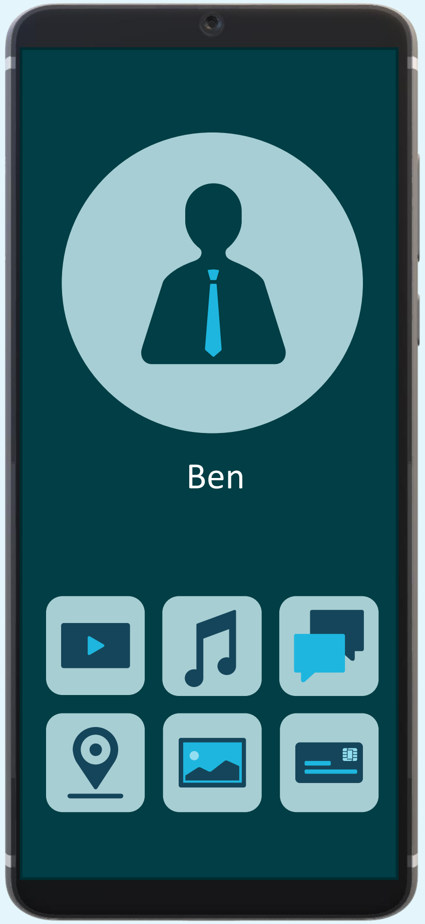
Personas and scenarios

Expert developer

Seasoned ML warrior

Need to squeeze all the performance offered by hardware



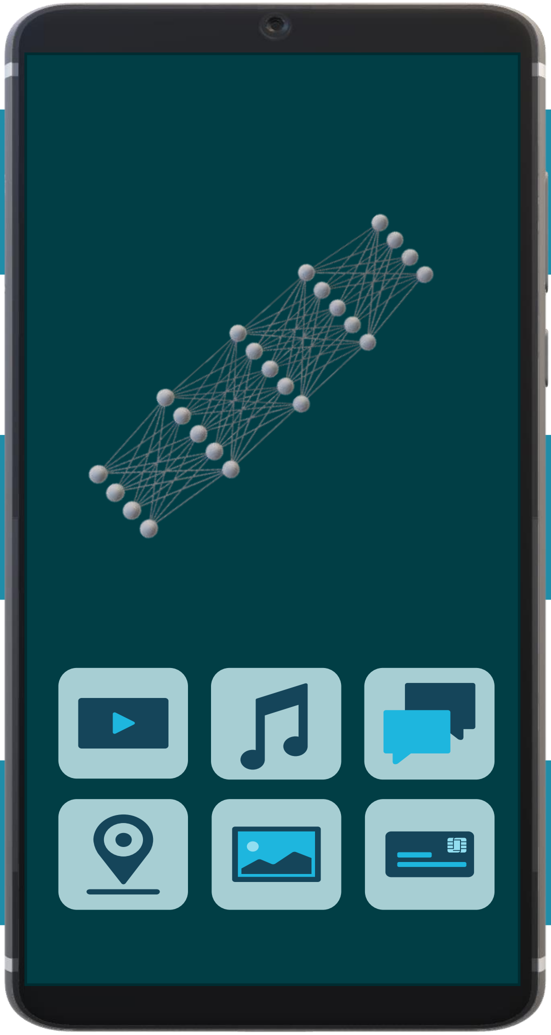


Novice developer

Not very sophisticated at ML

Scalability is more important than performance

Building an Application - Misconceptions



ML in the cloud and edge device is similar

Nothing could be further from reality

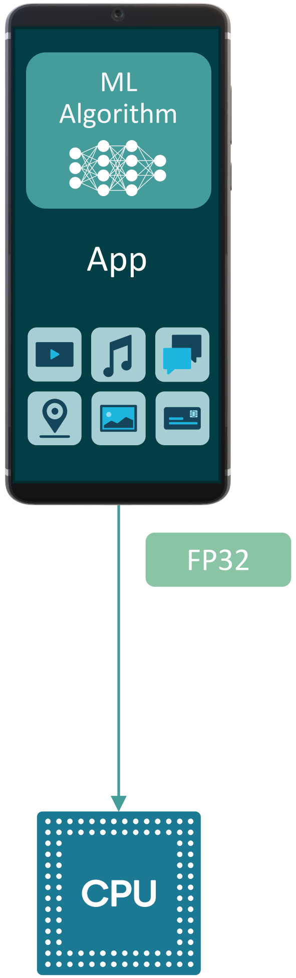
All runtime frameworks offer the same performance and flexibility

Runtime frameworks differ in cadence, range and performance

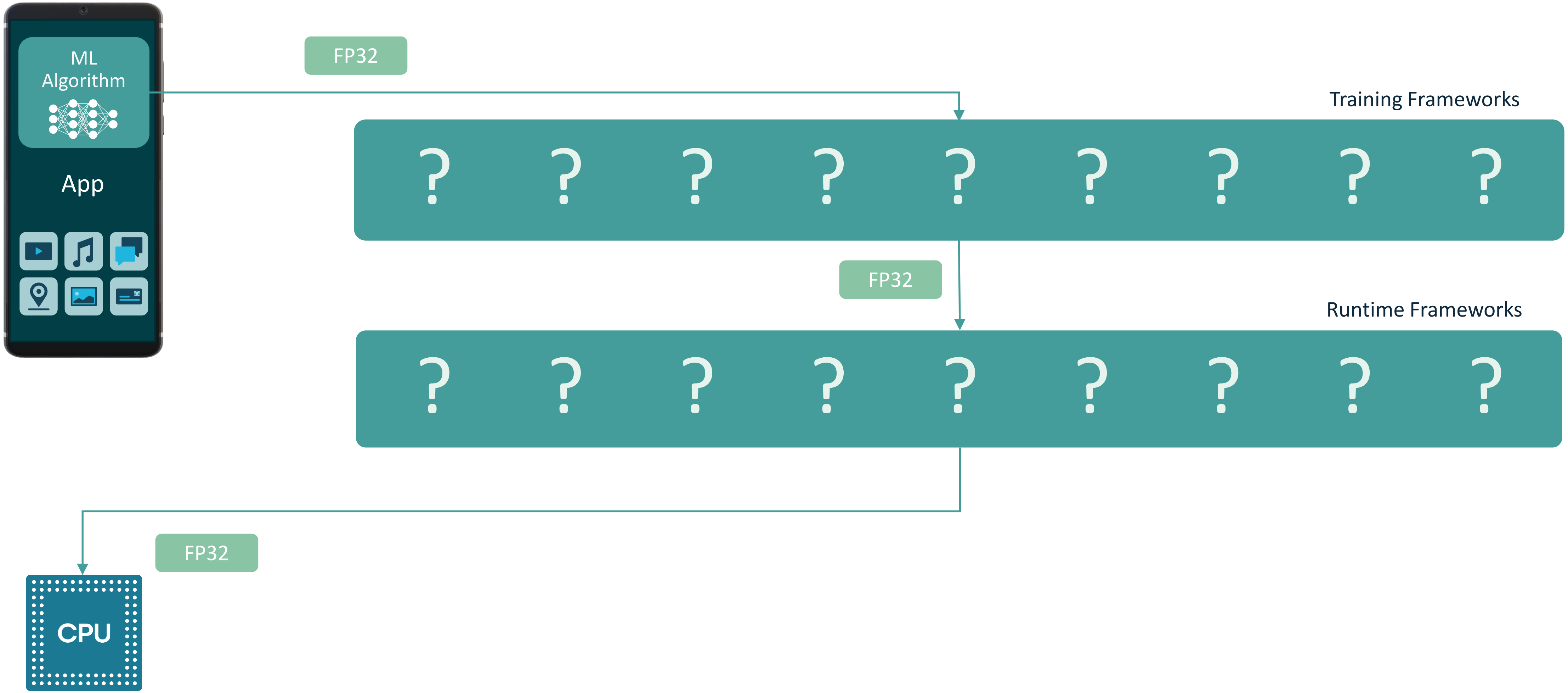
Quantization is hard and offers little benefit

Tools are available to offer users > 6 times in performance and power improvements

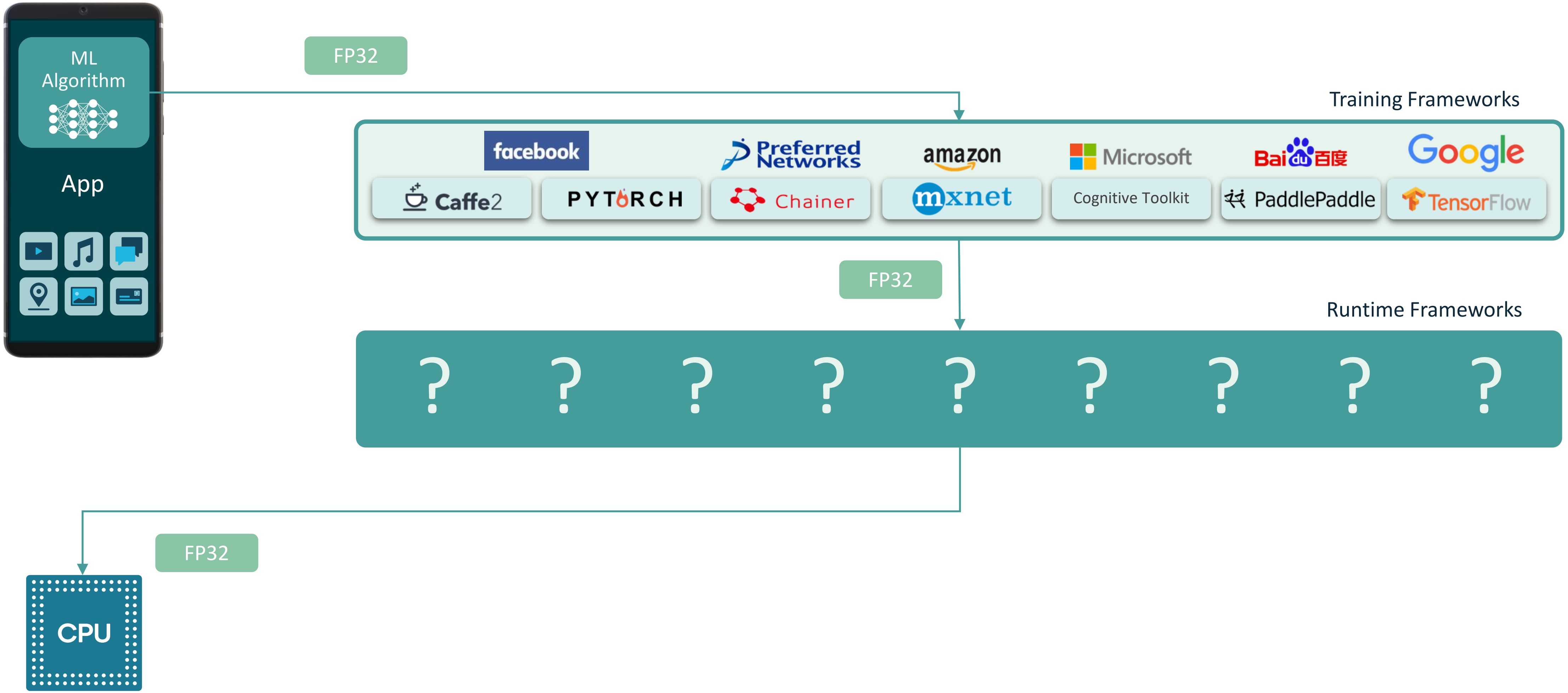
Building an Application



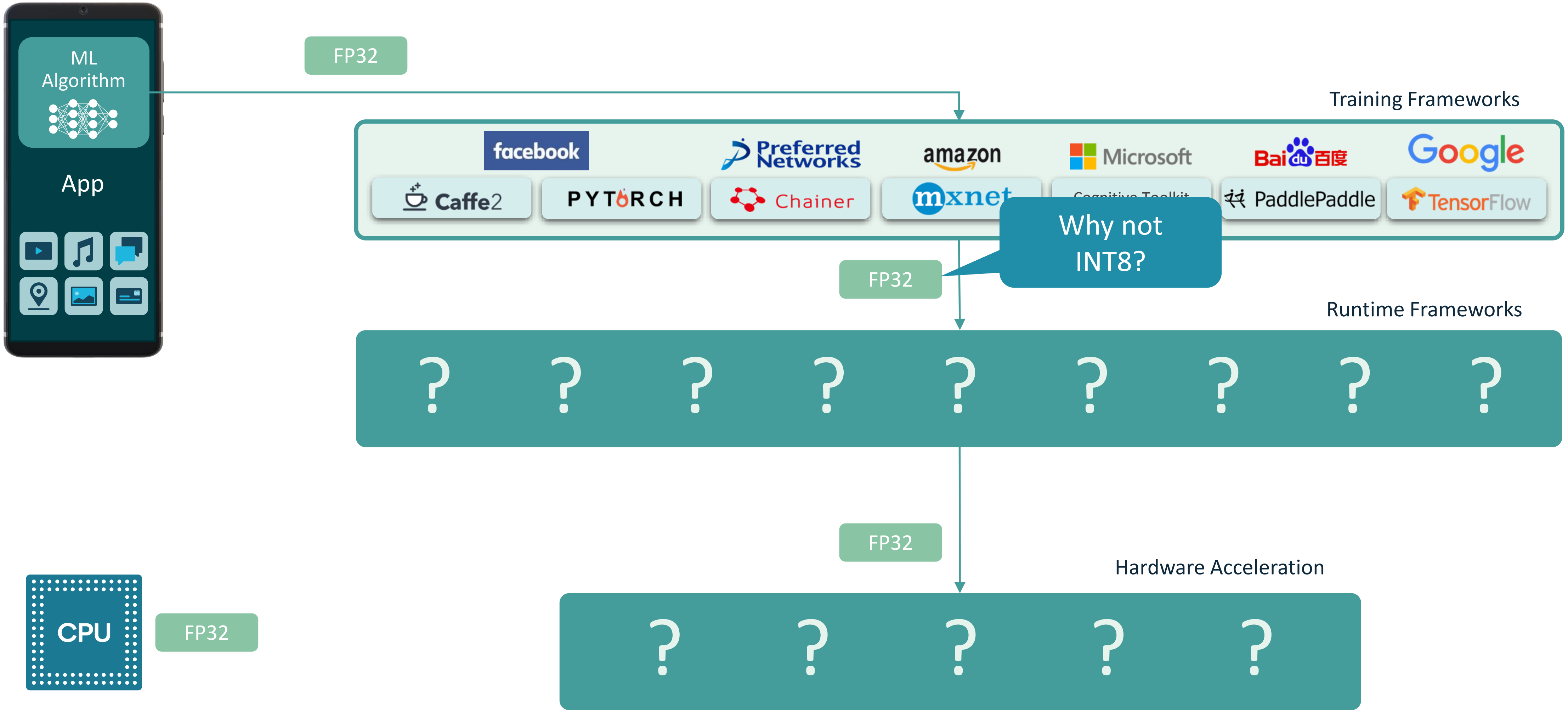
Adding ML to your Application



Adding ML to your Application

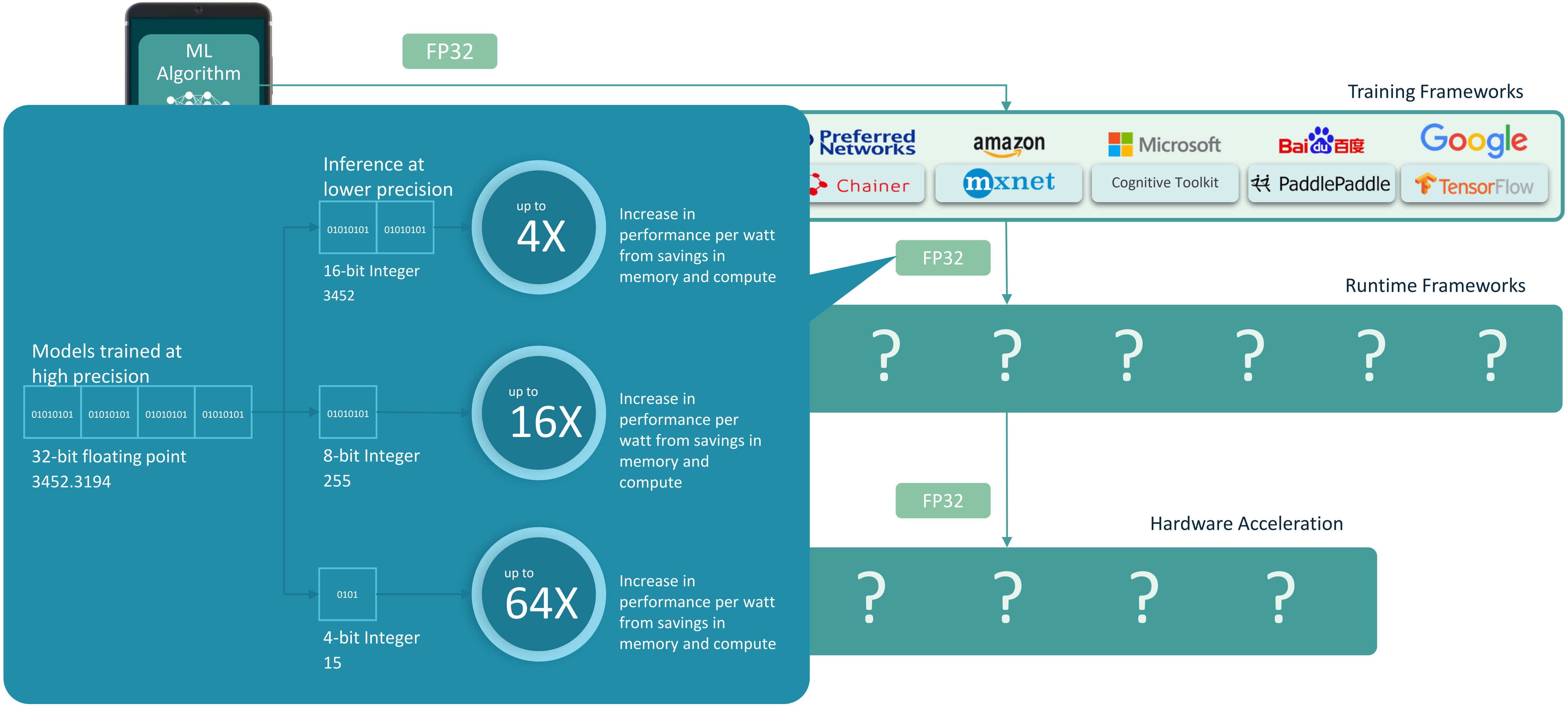


What if it is not accurate enough?



Why not INT8?

What if it is not accurate enough?



ML Algorithm

FP32

Training Frameworks

Preferred Networks — Chainer

amazon — mxnet

Microsoft — Cognitive Toolkit

Baidu 百度 — PaddlePaddle

Google — TensorFlow

FP32

Runtime Frameworks

? ? ? ? ? ?

FP32

Hardware Acceleration

? ? ? ?

Models trained at high precision

01010101 01010101 01010101 01010101

32-bit floating point

3452.3194

Inference at lower precision

01010101 01010101

16-bit Integer

3452

up to 4X

Increase in performance per watt from savings in memory and compute

01010101

8-bit Integer

255

up to 16X

Increase in performance per watt from savings in memory and compute

0101

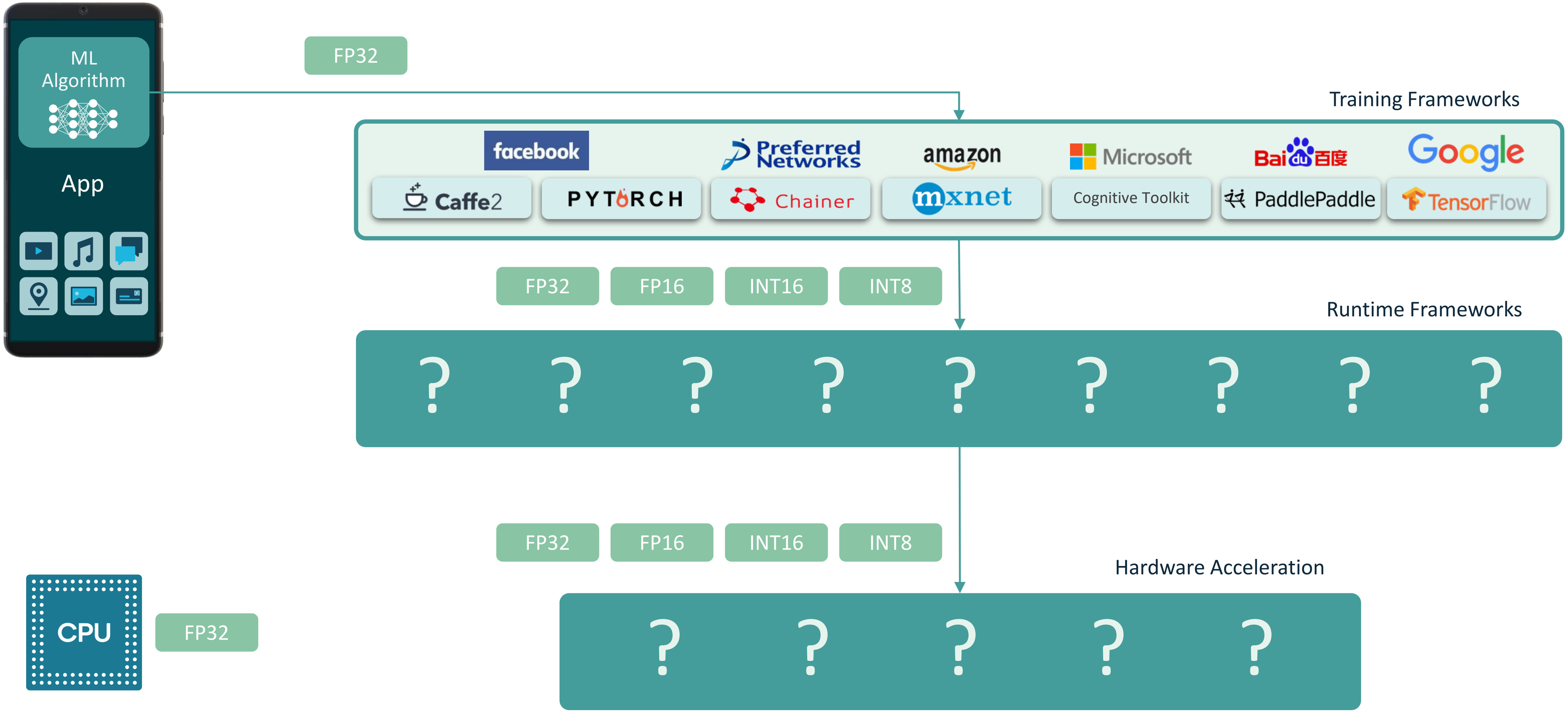
4-bit Integer

15

up to 64X

Increase in performance per watt from savings in memory and compute

Adding more data types



ML Algorithm

App

FP32

Training Frameworks

facebook

Preferred Networks

amazon

Microsoft

Bai百度

Google

Caffe2

PYTORCH

Chainer

mxnet

Cognitive Toolkit

PaddlePaddle

TensorFlow

FP32 FP16 INT16 INT8

Runtime Frameworks

? ? ? ? ? ? ? ? ?

FP32 FP16 INT16 INT8

Hardware Acceleration

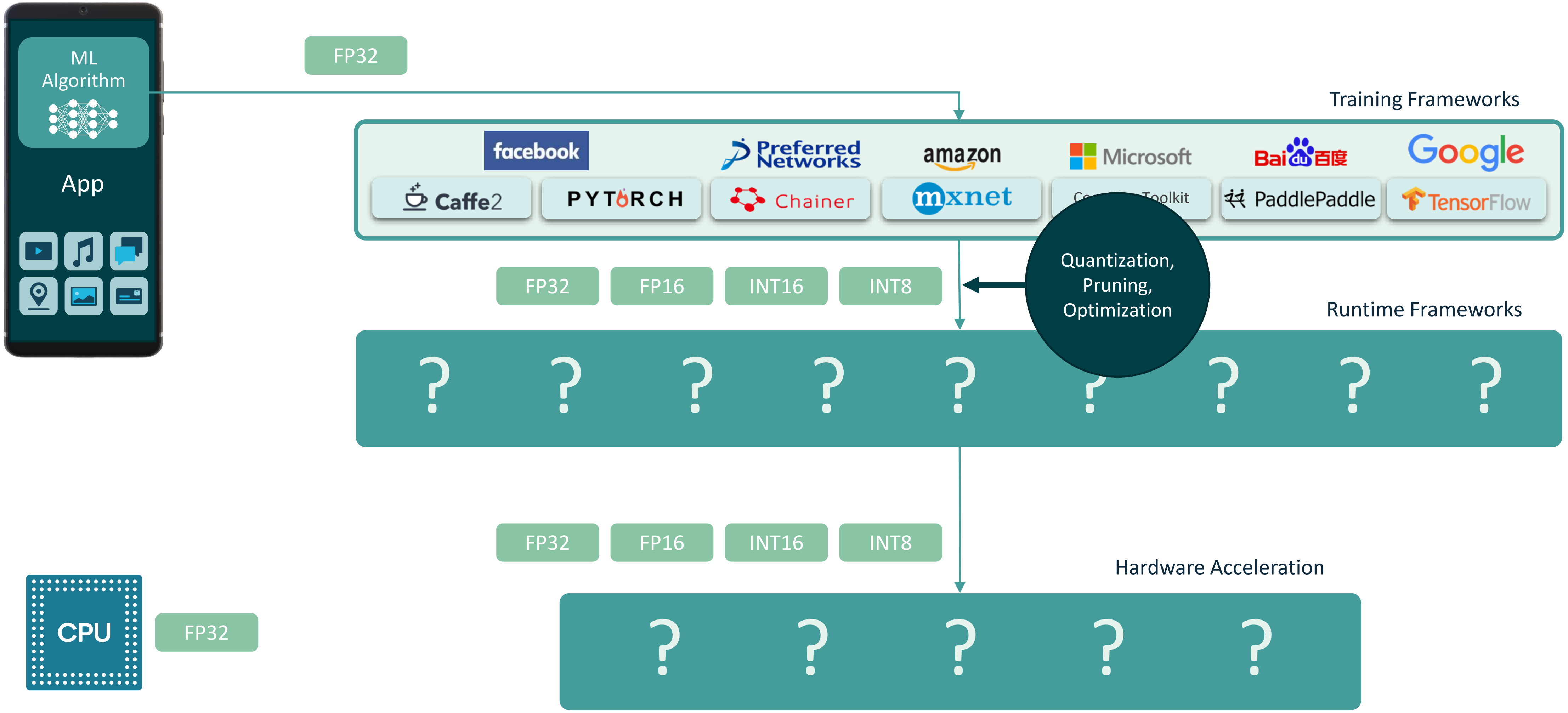
? ? ? ? ?

CPU

FP32

Qualcomm

Adding more data types – need tools to make it seamless



ML Algorithm

App

FP32

Training Frameworks

facebook

Preferred Networks

amazon

Microsoft

Baidu 百度

Google

Caffe2

PYTORCH

Chainer

mxnet

Co Toolkit

PaddlePaddle

TensorFlow

Quantization, Pruning, Optimization

FP32 FP16 INT16 INT8

Runtime Frameworks

? ? ? ? ? ? ? ? ?

FP32 FP16 INT16 INT8

Hardware Acceleration

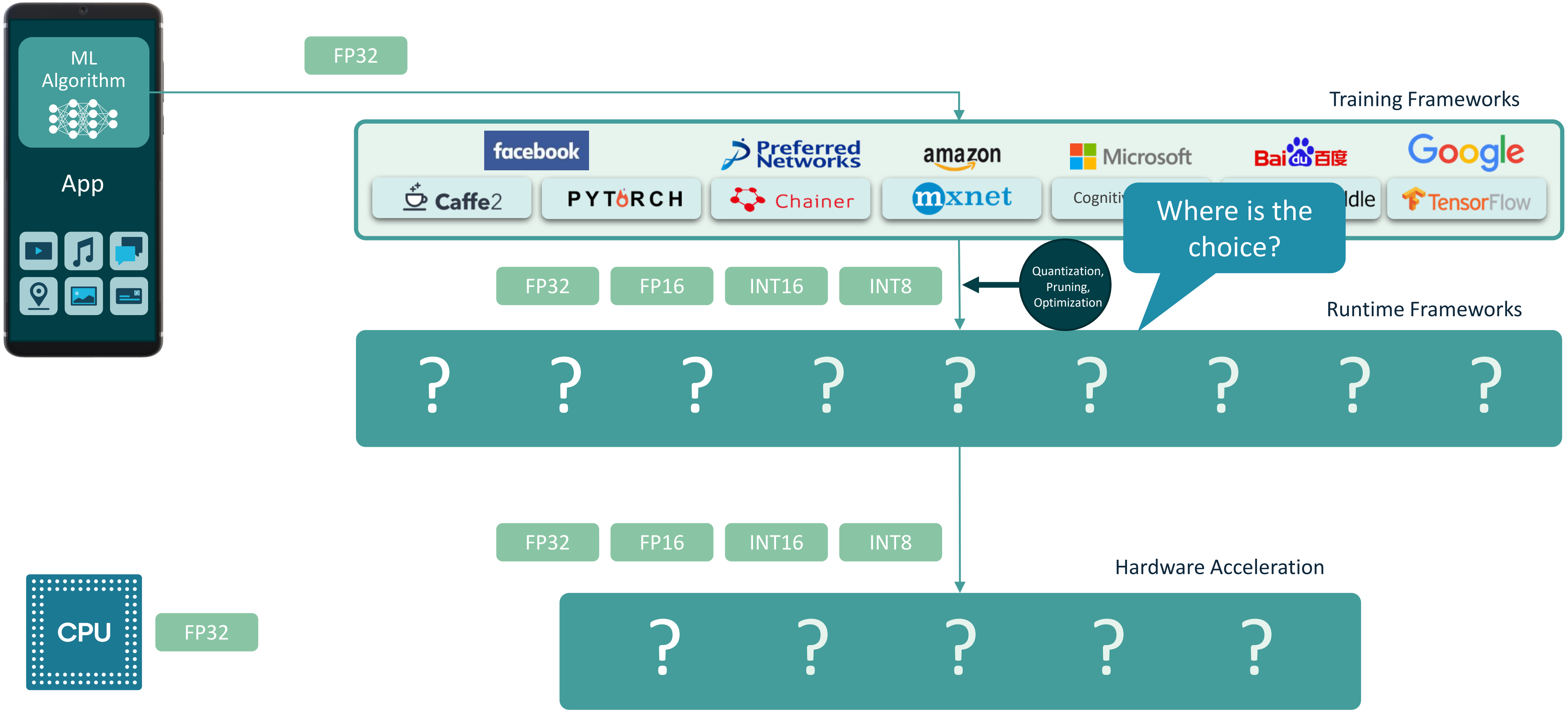
? ? ? ? ?

CPU

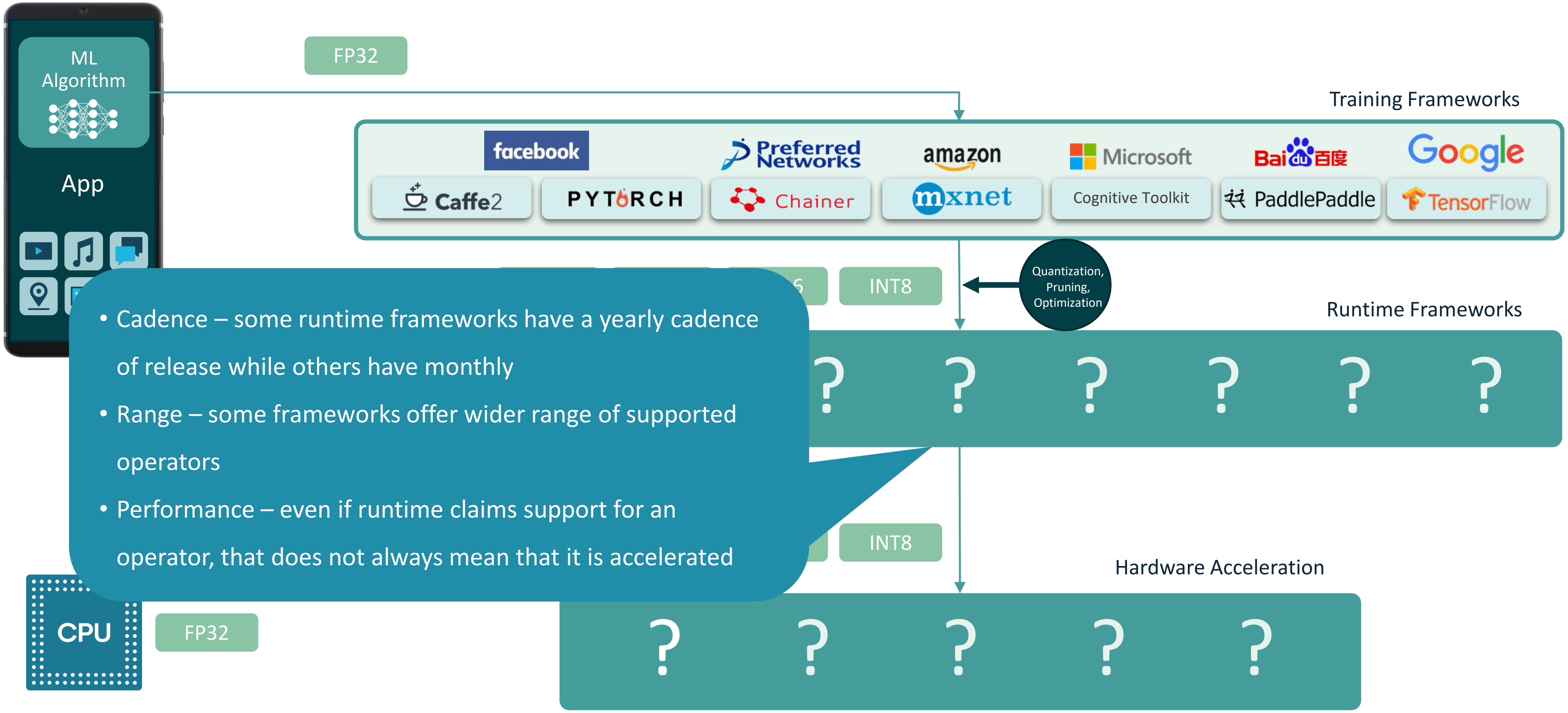
FP32

Qualcomm

What if it is not flexible enough?

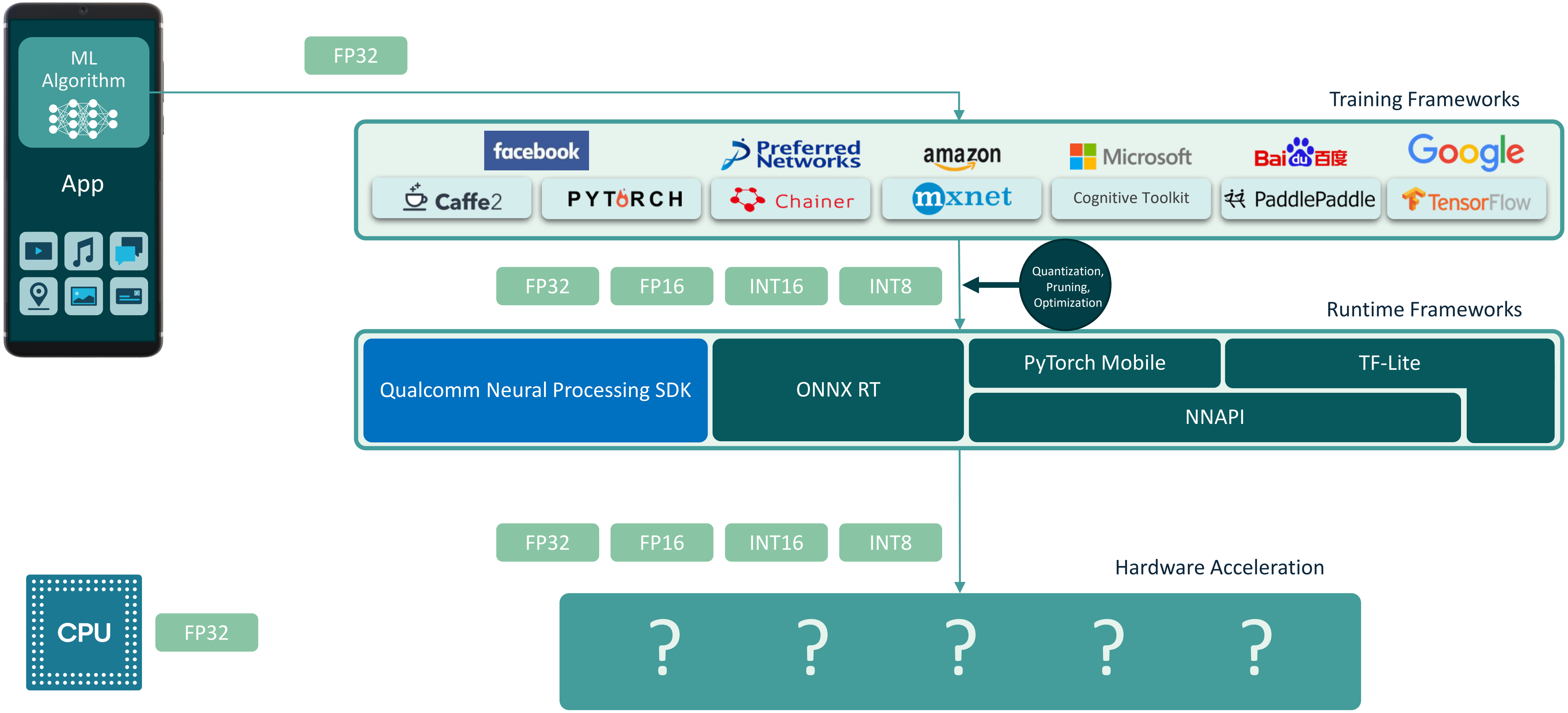


What if it is not flexible enough?



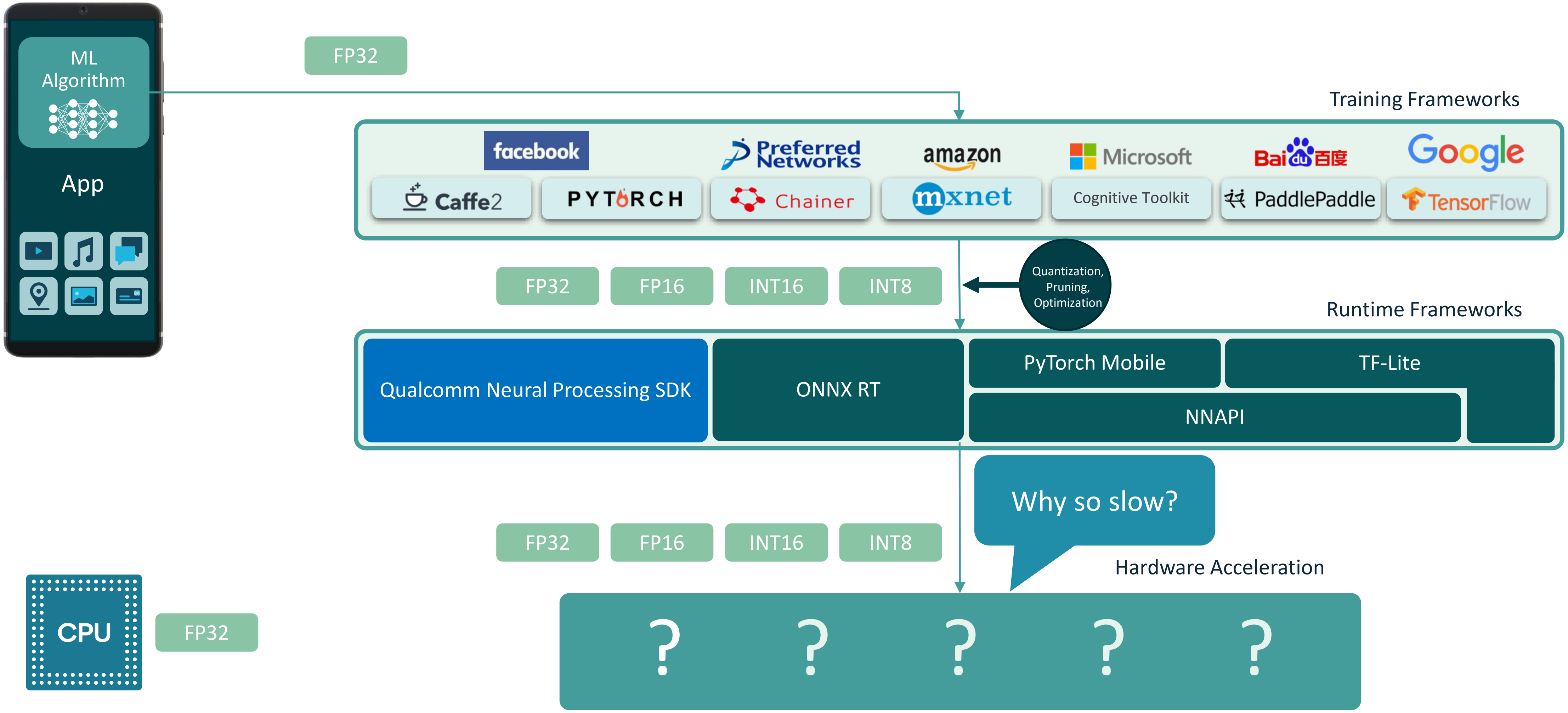
- Cadence – some runtime frameworks have a yearly cadence of release while others have monthly
- Range – some frameworks offer wider range of supported operators
- Performance – even if runtime claims support for an operator, that does not always mean that it is accelerated

Selecting a runtime framework that fits



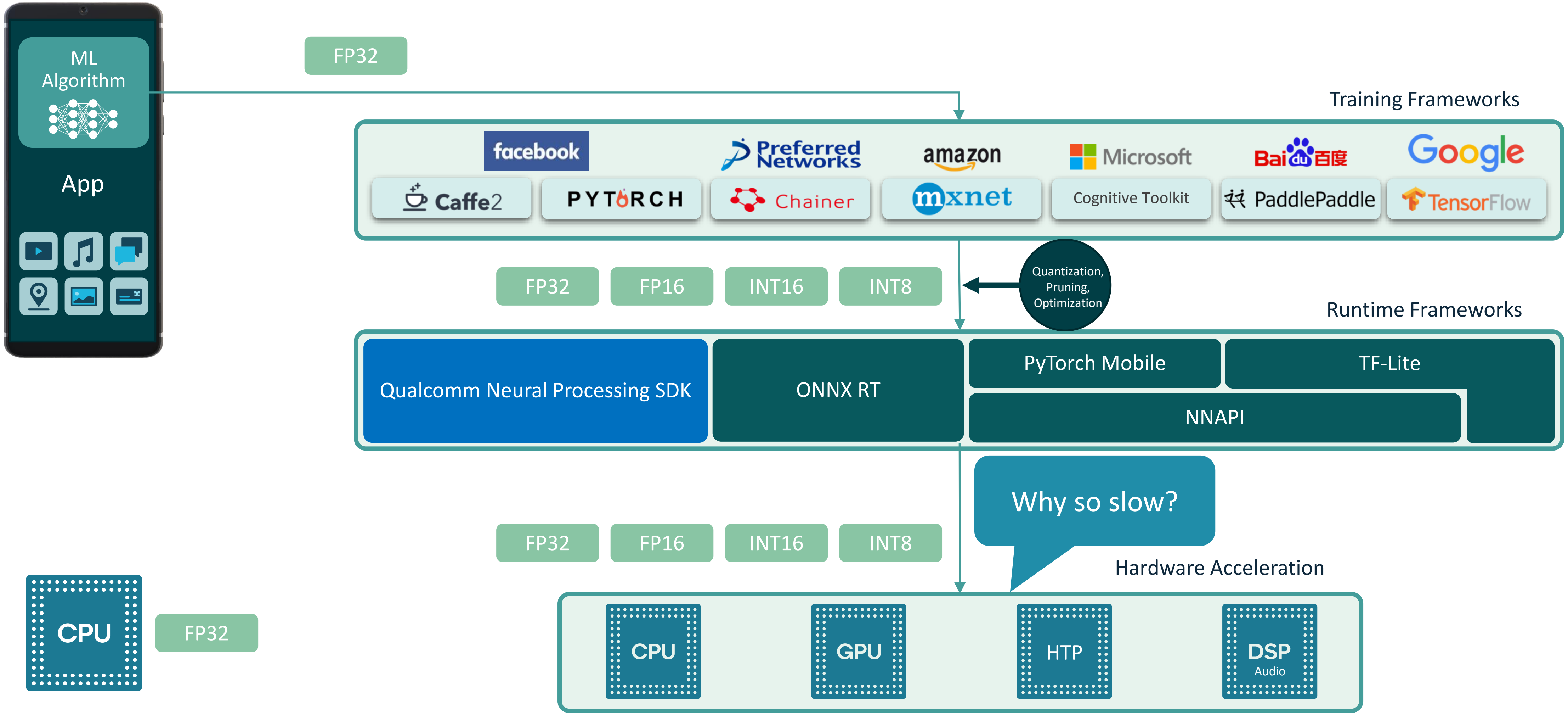
Qualcomm Neural Processing SDK is a product of Qualcomm Technologies, Inc and/or its subsidiaries.

What if it is not flexible enough?



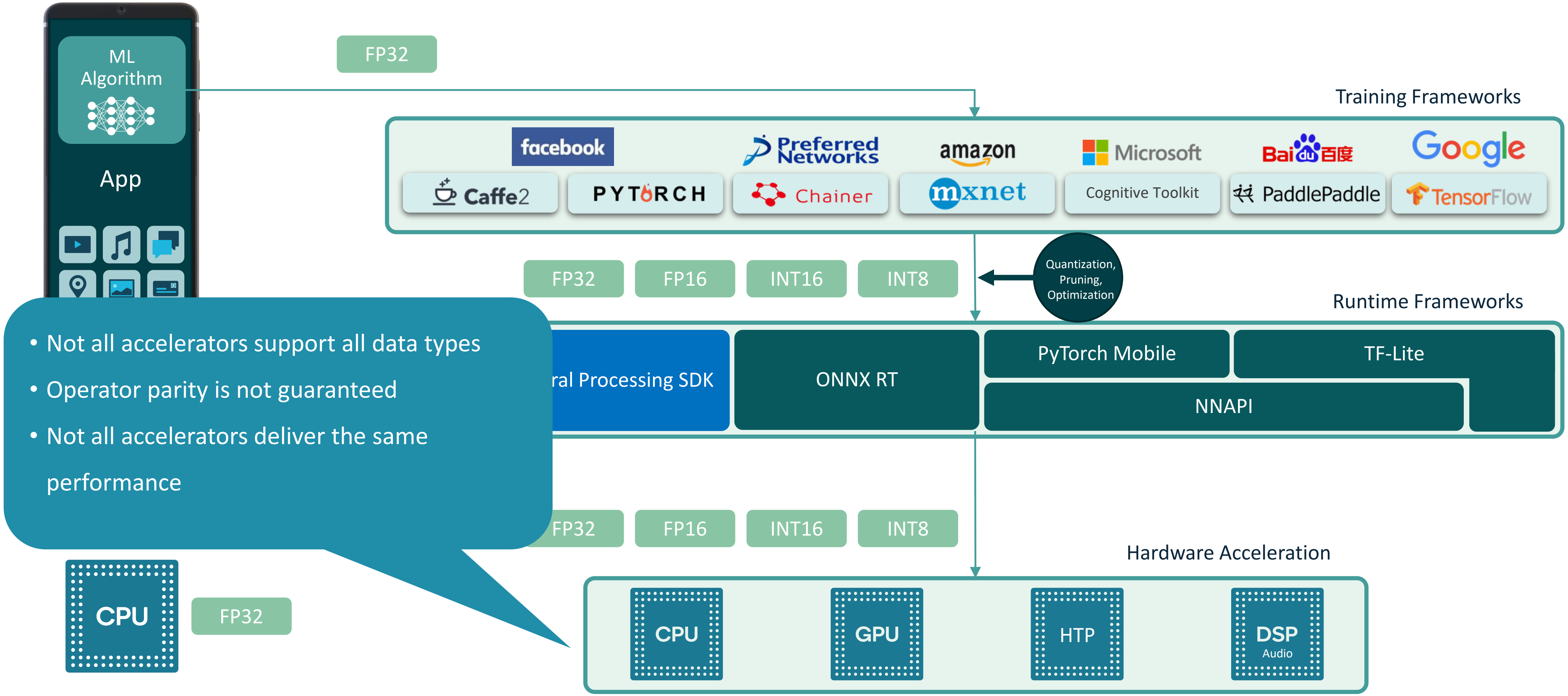
Qualcomm Neural Processing SDK is a product of Qualcomm Technologies, Inc and/or its subsidiaries.

What if it is not fast enough?



Qualcomm Neural Processing SDK is a product of Qualcomm Technologies, Inc and/or its subsidiaries.

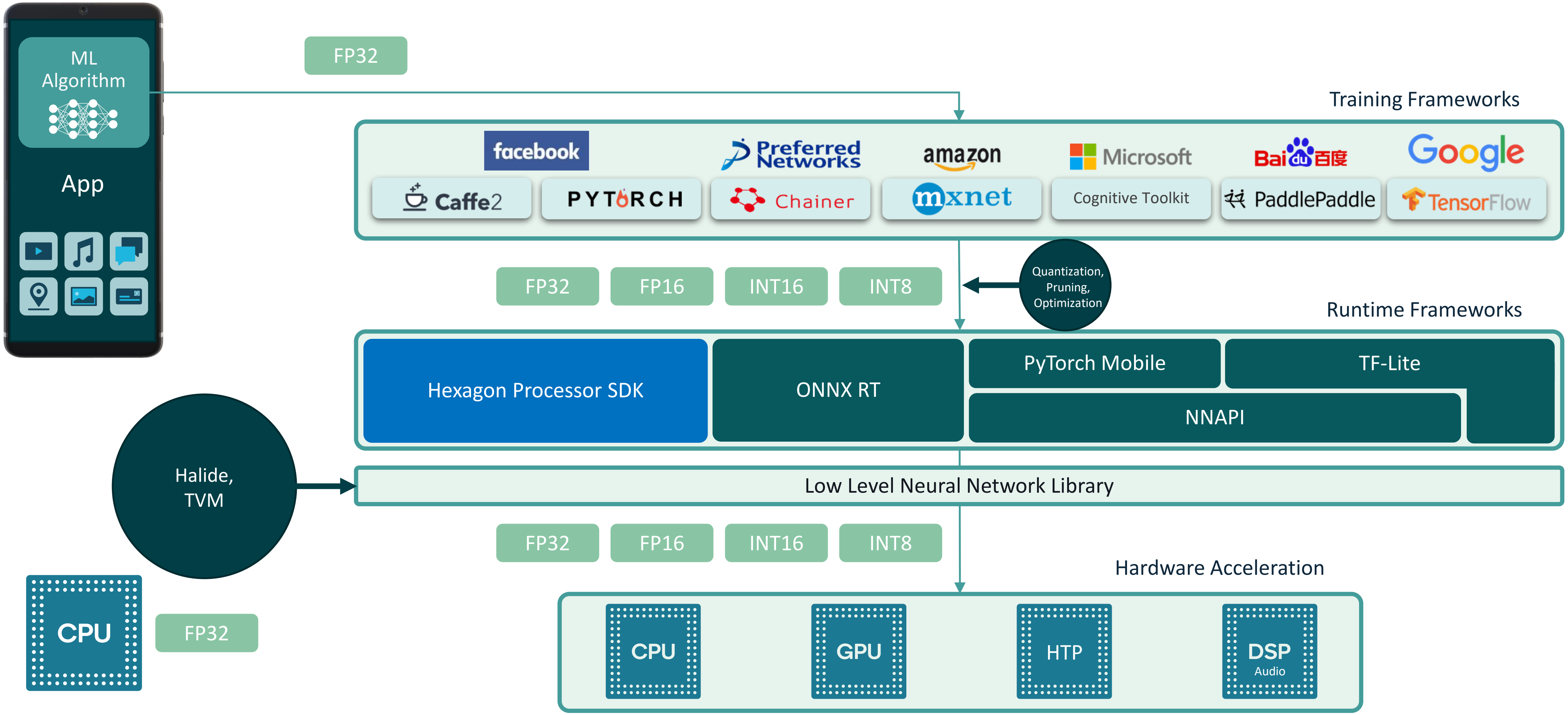
What if it is not fast enough?



- Not all accelerators support all data types
- Operator parity is not guaranteed
- Not all accelerators deliver the same performance

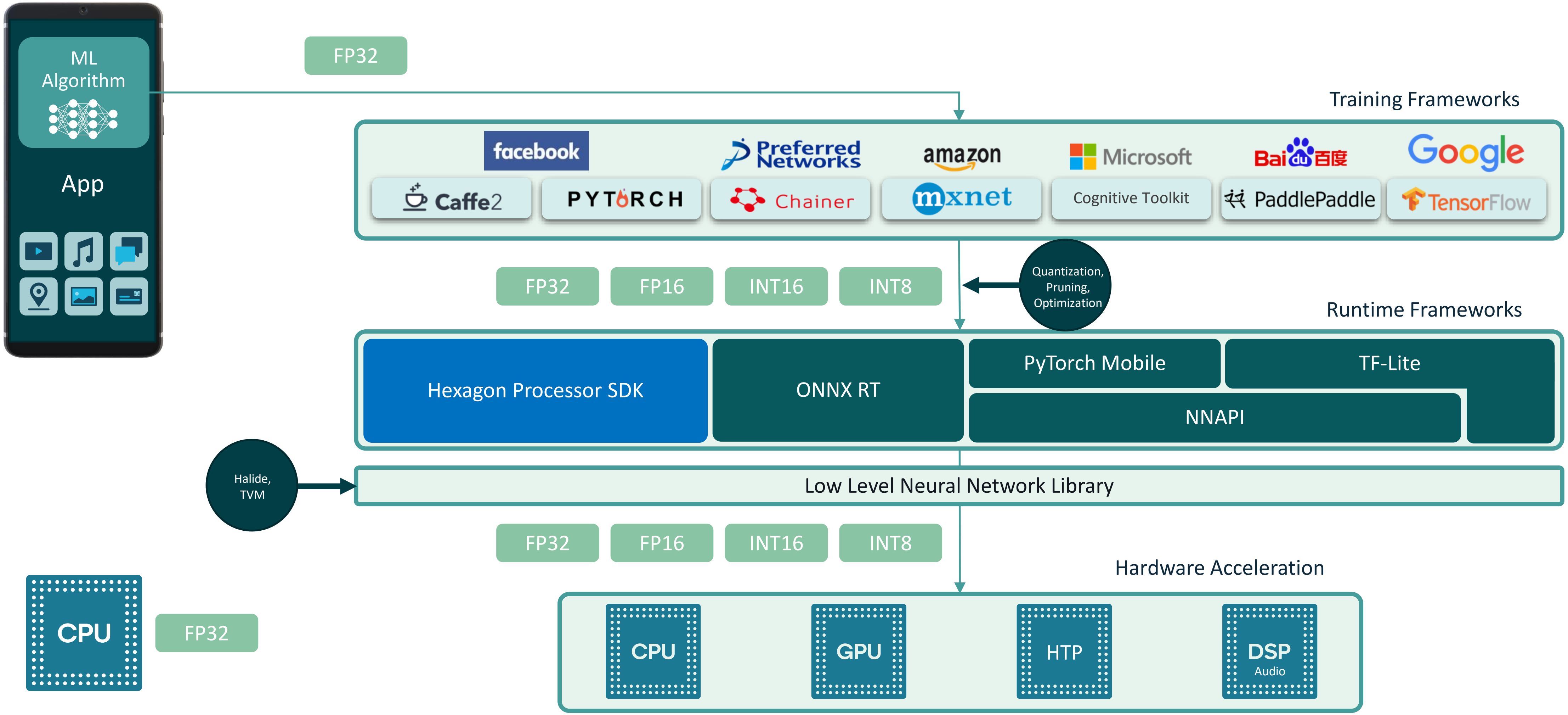
Qualcomm Neural Processing SDK is a product of Qualcomm Technologies, Inc and/or its subsidiaries.

Selecting a runtime framework that fits



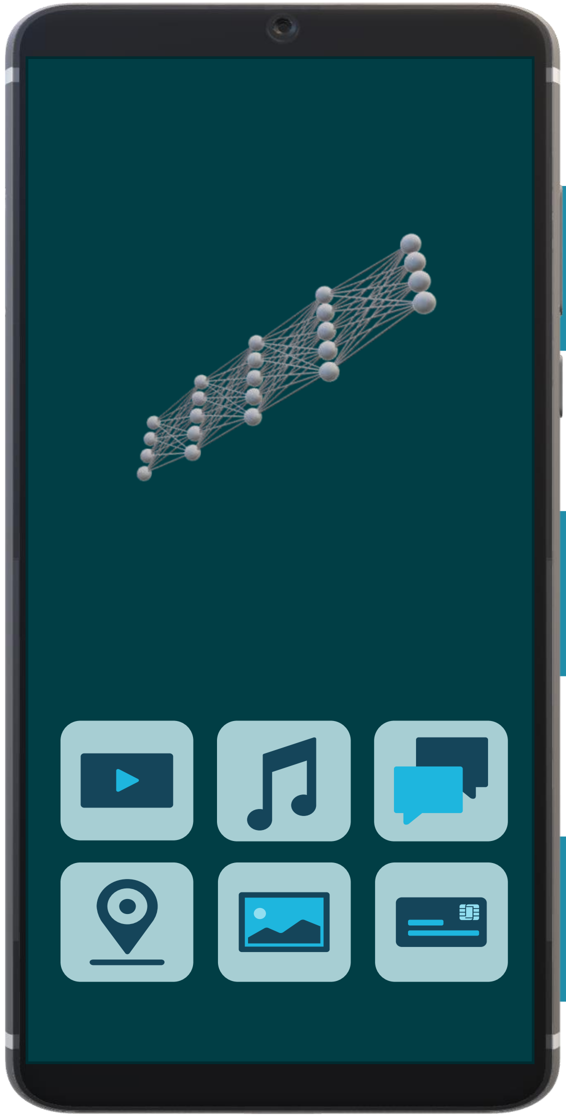
Qualcomm Neural Processing SDK is a product of Qualcomm Technologies, Inc and/or its subsidiaries.

Selecting a runtime framework that fits



Qualcomm Neural Processing SDK is a product of Qualcomm Technologies, Inc and/or its subsidiaries.

Key Takeaways



Not all applications are built the same way, your software stack will determine how well your application will perform

In order to achieve your application full capacity, you need a software stack that is tailored to specifically to what you are looking to accomplish

Different models require specific tools that only customizable stacks will offer

Qualcomm and Hexagon are trademarks or registered trademarks of Qualcomm Incorporated

2021 embedded VISION summit

VIRTUAL | MAY 25-28

Thank You

Qualcomm

Resource Slide

- **Qualcomm AI page:**

https://www.qualcomm.com/invention/artificial-intelligence

- **Qualcomm AI research:**

https://www.qualcomm.com/invention/artificial-intelligence/ai-research?cmpid=fofyus193556&gclid=CjwKCAjw19z6BRAYEiwAmo64LfQjU8vqH8TxqKTM2PZQp8JibXrjev85wLfKFknJnS_b494yZ7e_WhoCPQkQAvD_BwE

- **Qualcomm Platform Solution Ecosystem:**

https://www.qualcomm.com/support/qan/platform-solutions-ecosystem

- **GitHub AI Model Efficiency Toolkit (AIMET):**

https://github.com/quic/aimet

- **Qualcomm Mobile AI page:**

https://www.qualcomm.com/products/smartphones/mobile-ai

- **Qualcomm Mobile AI blog:**

https://www.qualcomm.com/news/onq/2020/12/02/exploring-ai-capabilities-qualcomm-snapdragon-888-mobile-platform

- **Qualcomm Cloud AI 100 blog:**

https://www.qualcomm.com/news/onq/2021/03/15/qualcomm-cloud-ai-100-amd-epyc-7003-series-processor-and-gigabyte-server