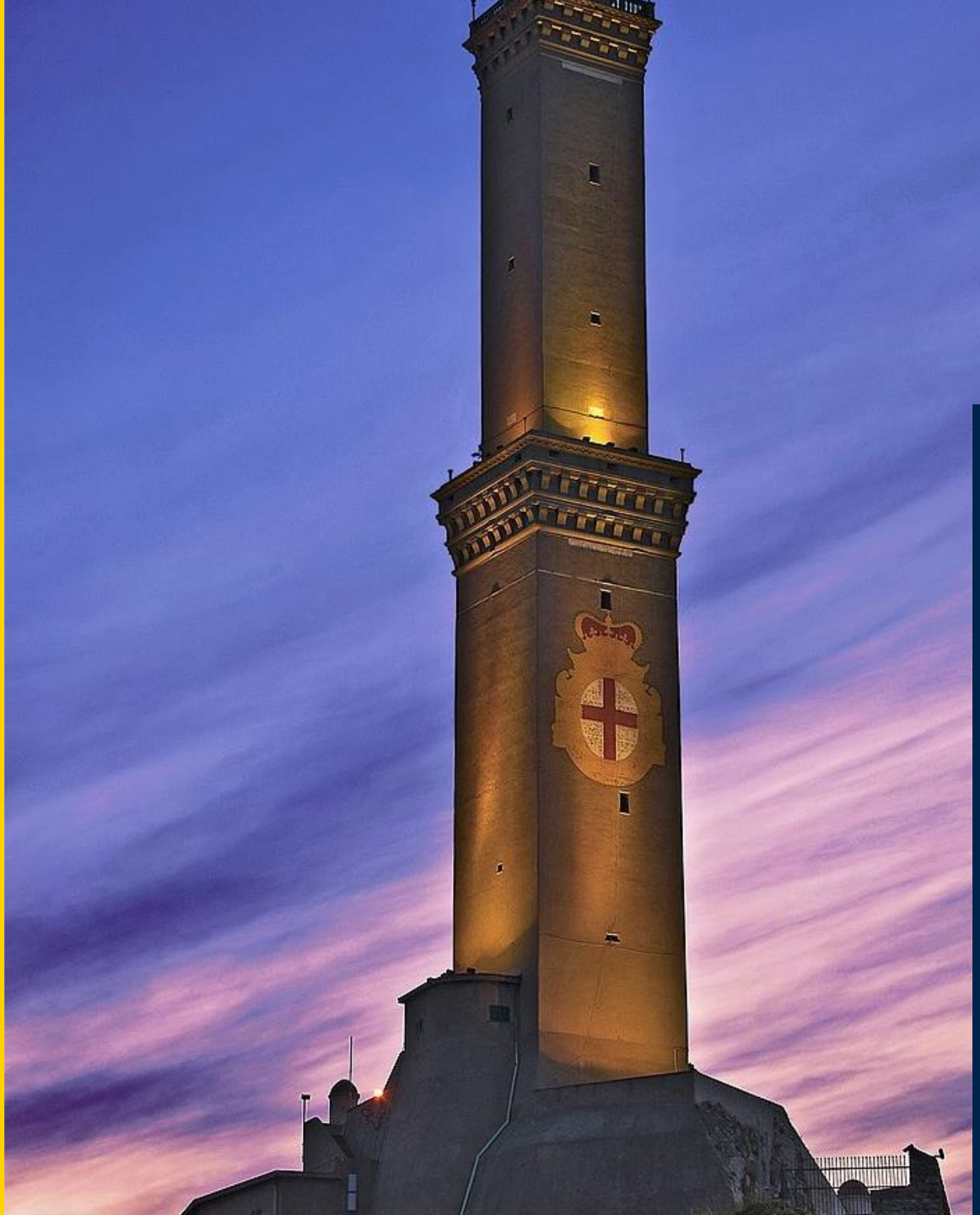


life.augmented

From Heterogenous Development Frameworks to Unified Benchmarking in the Cloud

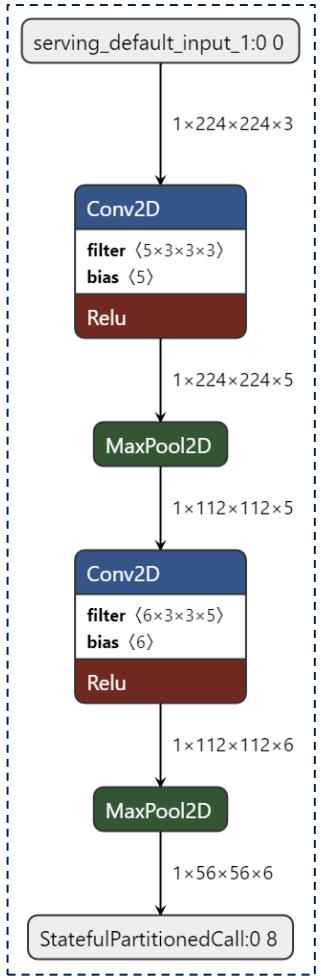
Marco Lattuada, PhD

STMicroelectronics

Introduction

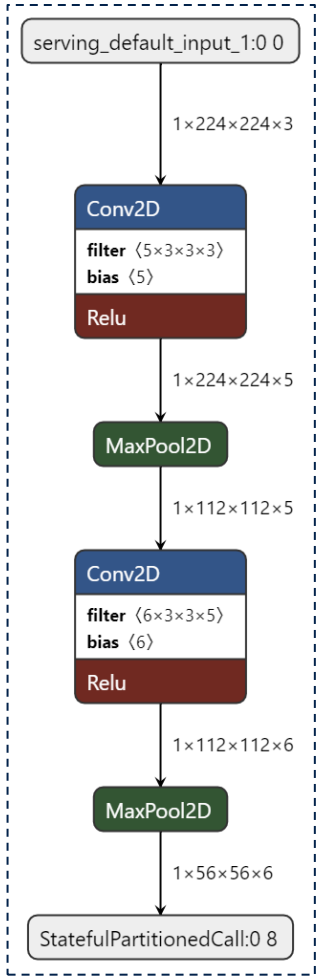
- Deployment of **Machine Learning** models on **Embedded Systems** is becoming **more and more relevant**
- Machine learning models can be very different in terms of algorithms, size, formats, etc.
- An automatic deployment flow must tackle this **heterogeneity** to reduce as much as possible the restrictions to the end users
- This presentation will show:
  - What characterize heterogeneity of Machine Learning models
  - How heterogeneity can be addressed to **maximize the support** and to **maximize sharing and reuse**

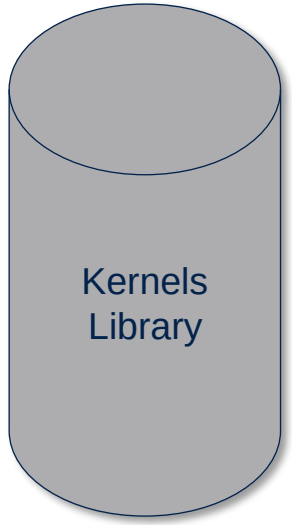
From Machine Learning to Implementation



life.augmented

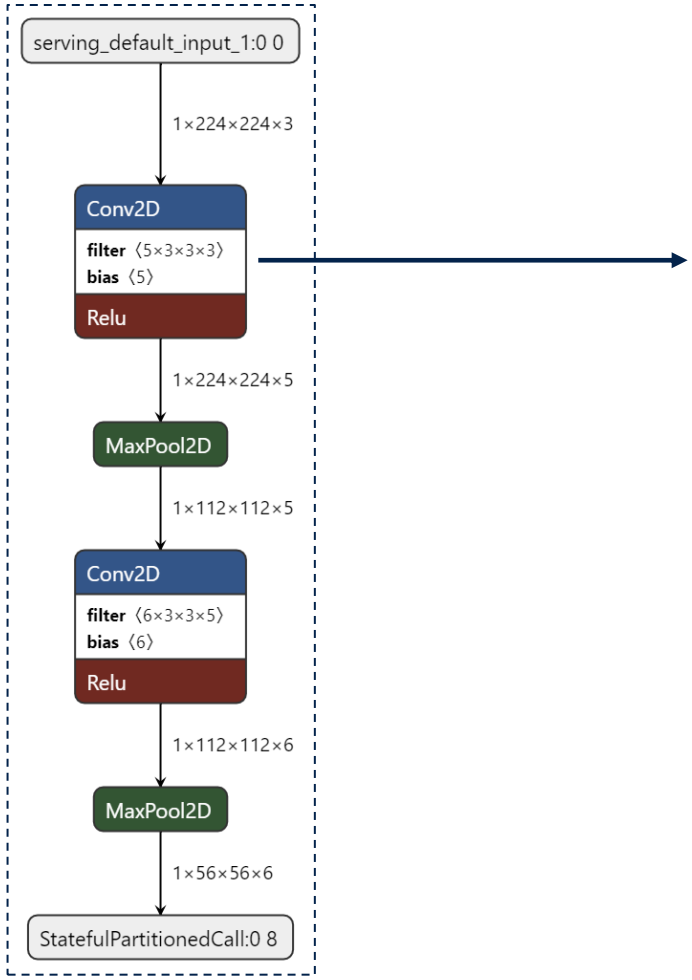
From Machine Learning to Implementation



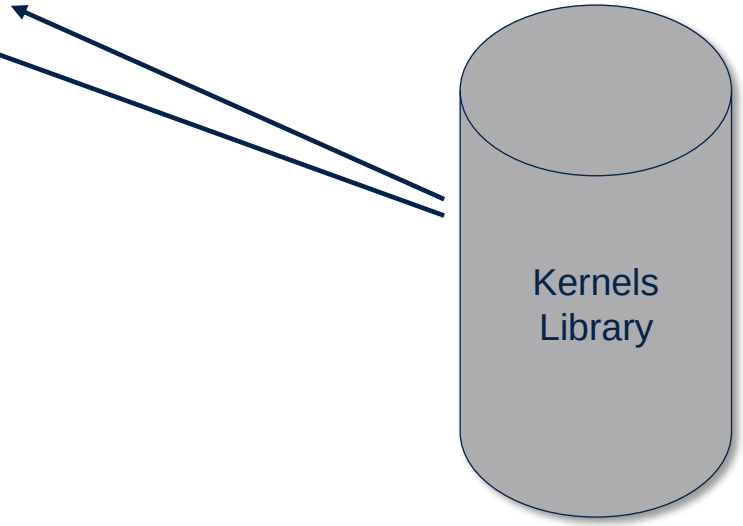


life.augmented

From Machine Learning to Implementation

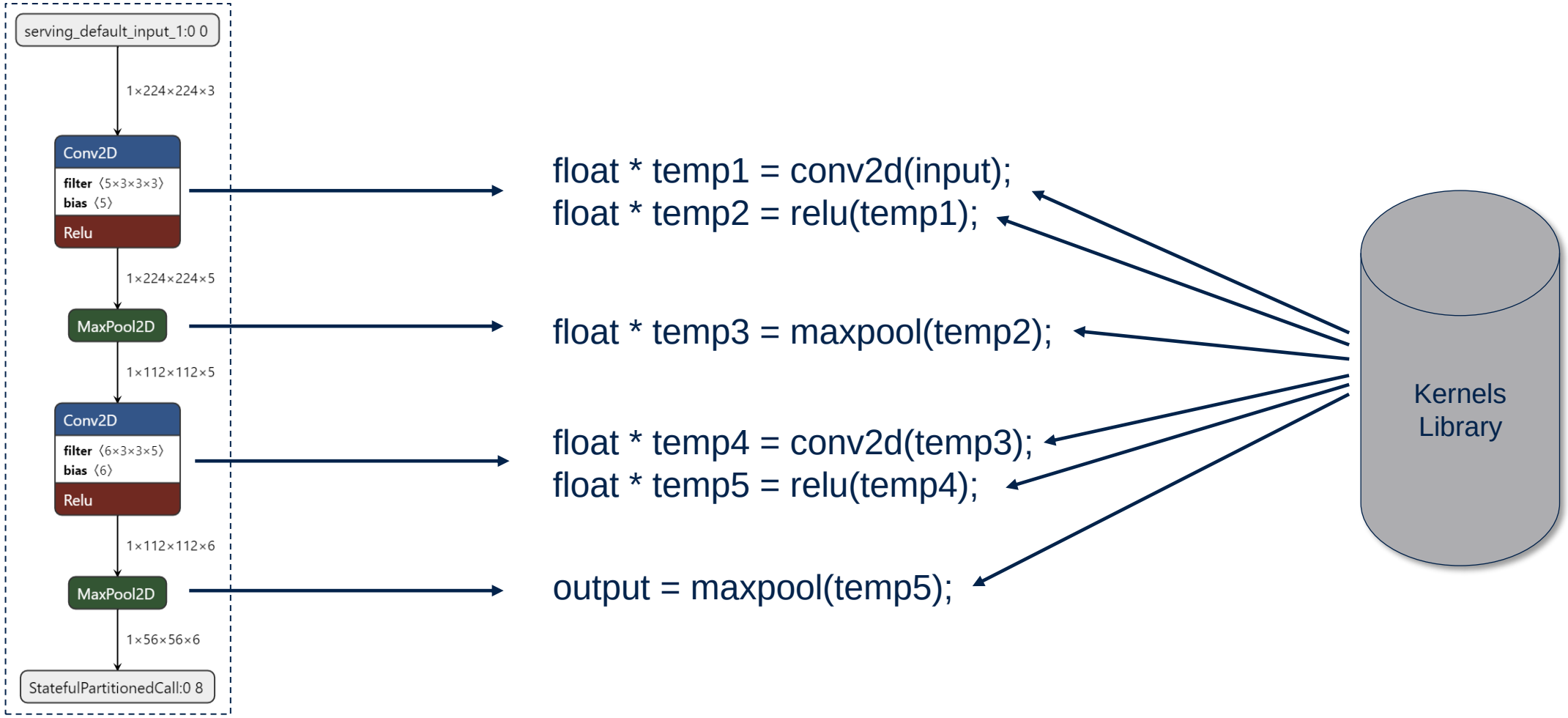



```
float * temp1 = conv2d(input);
float * temp2 = relu(temp1);
```

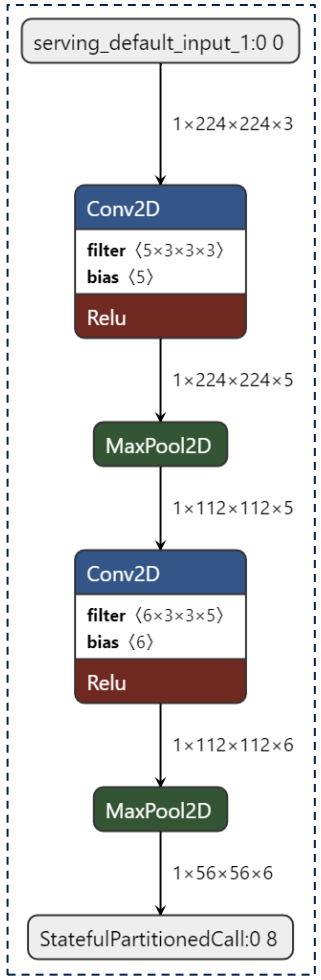


life.augmented

From Machine Learning to Implementation



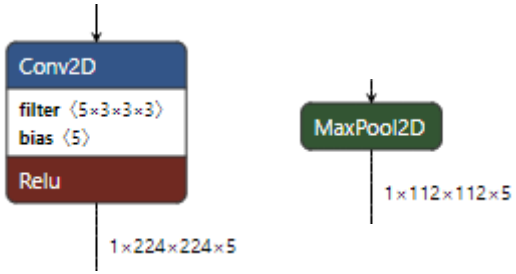
From Machine Learning to Implementation




```
void network(float * input, float * output) {

    float * temp1 = conv2d(input);
    float * temp2 = relu(temp1);

    float * temp3 = maxpool(temp2);

    float * temp4 = conv2d(temp3);
    float * temp5 = relu(temp4);

    output = maxpool(temp5);

}
```

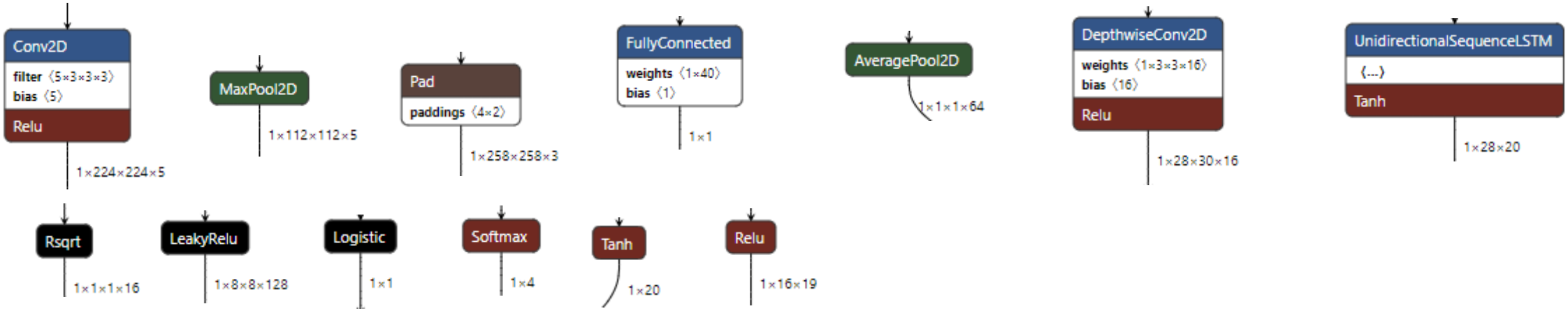
Kernels Library

Heterogeneity: Operators

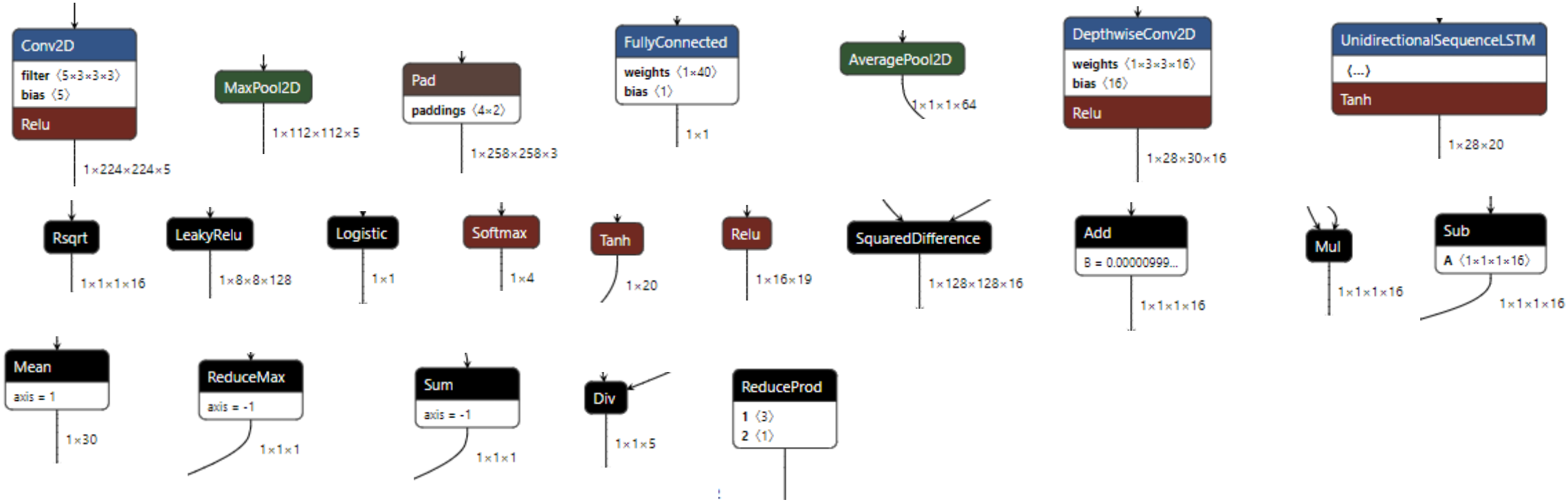


Heterogeneity: Operators

Heterogeneity: Operators

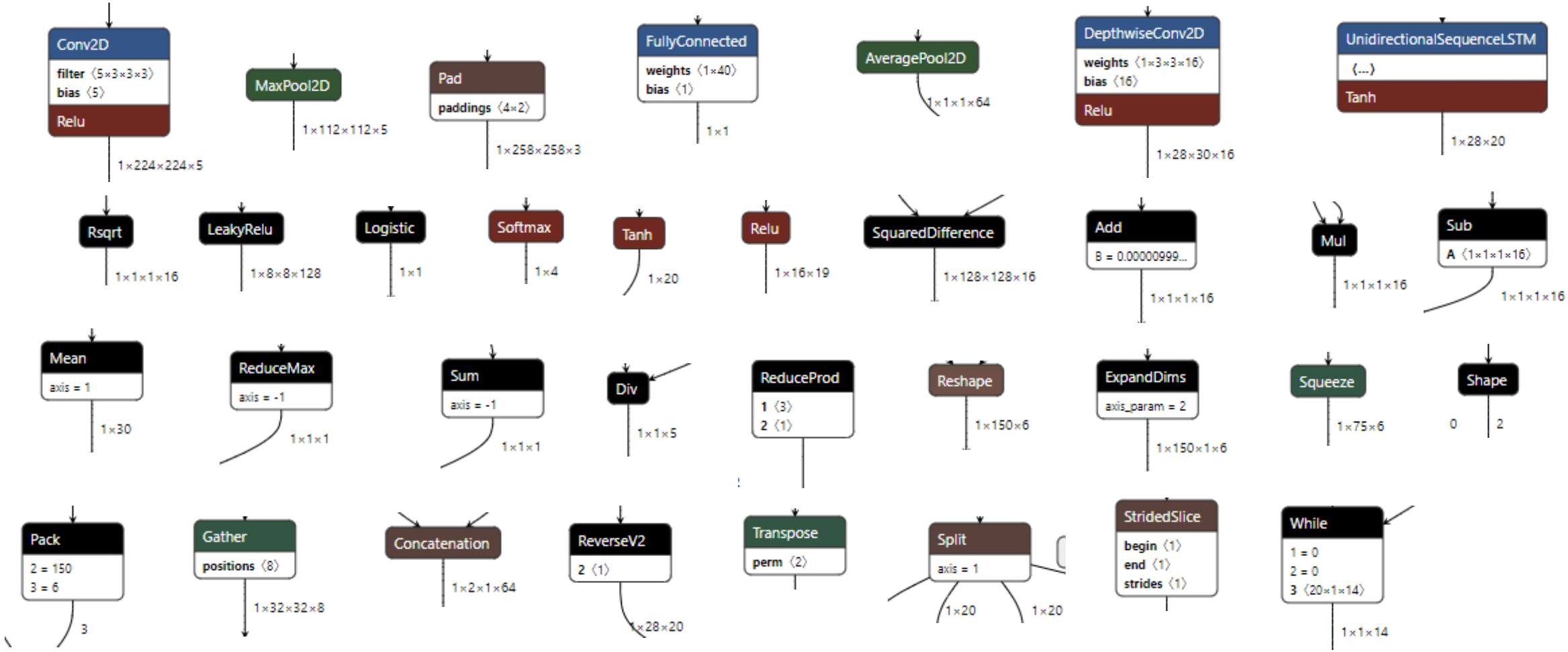


Heterogeneity: Operators



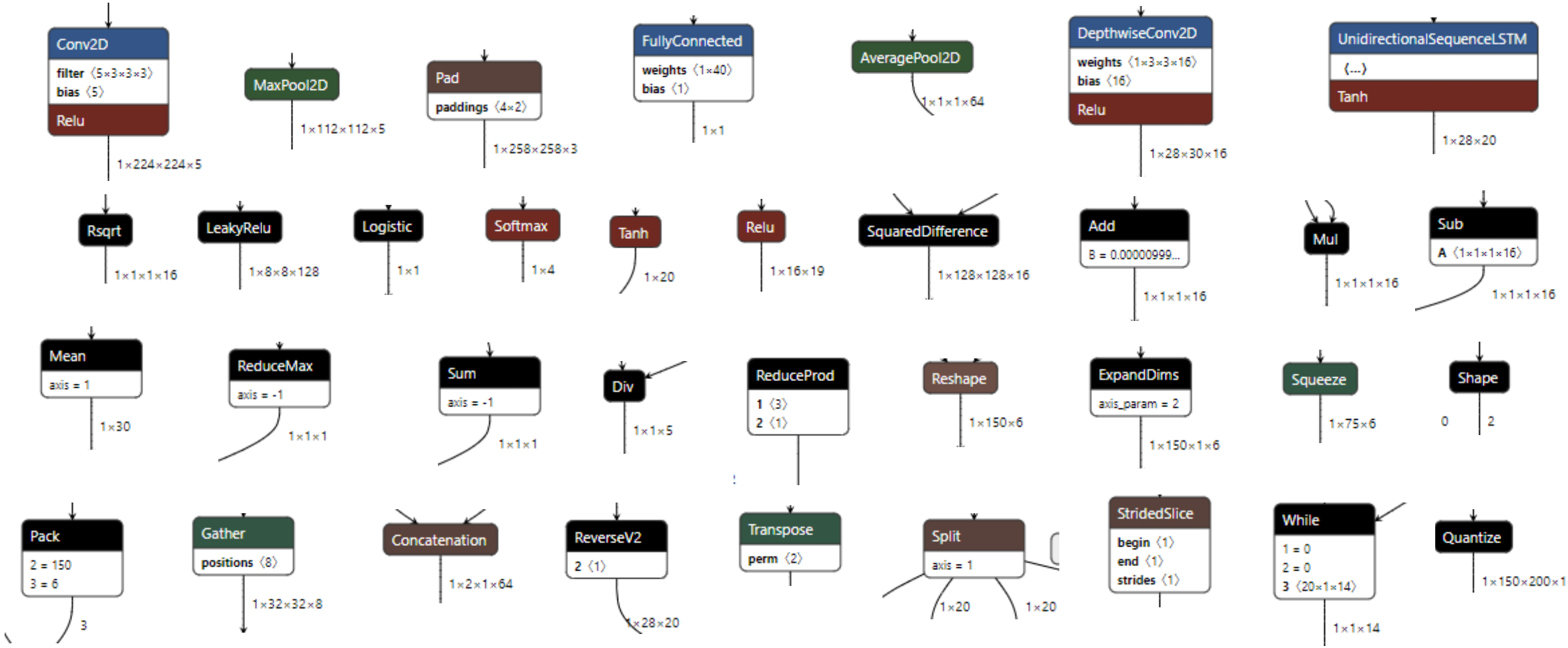
life.augmented

Heterogeneity: Operators



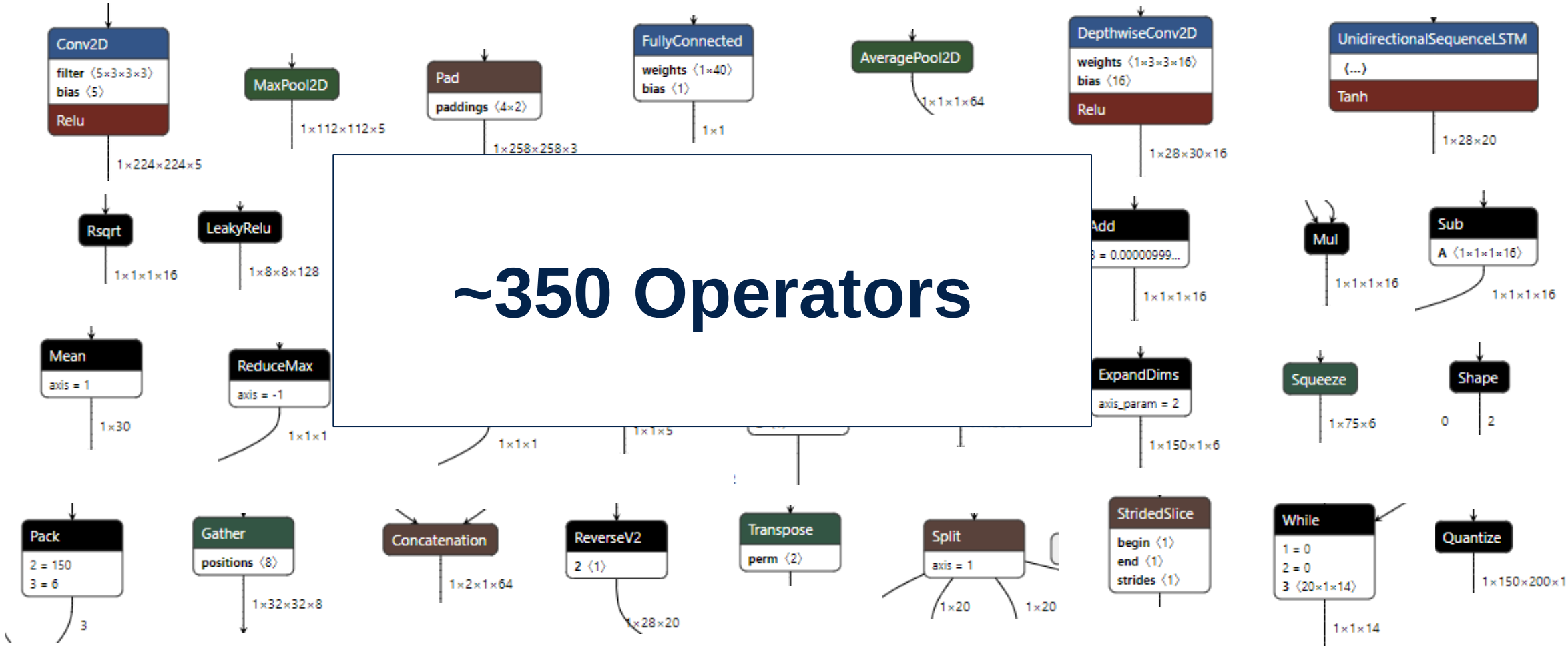
life.augmented

Heterogeneity: Operators



life.augmented

Heterogeneity: Operators

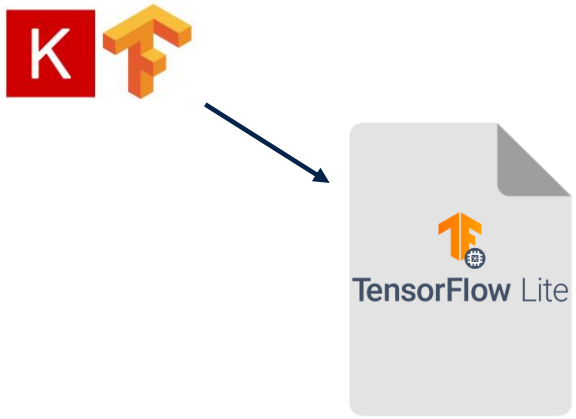


~350 Operators

Heterogeneity: DL Formats

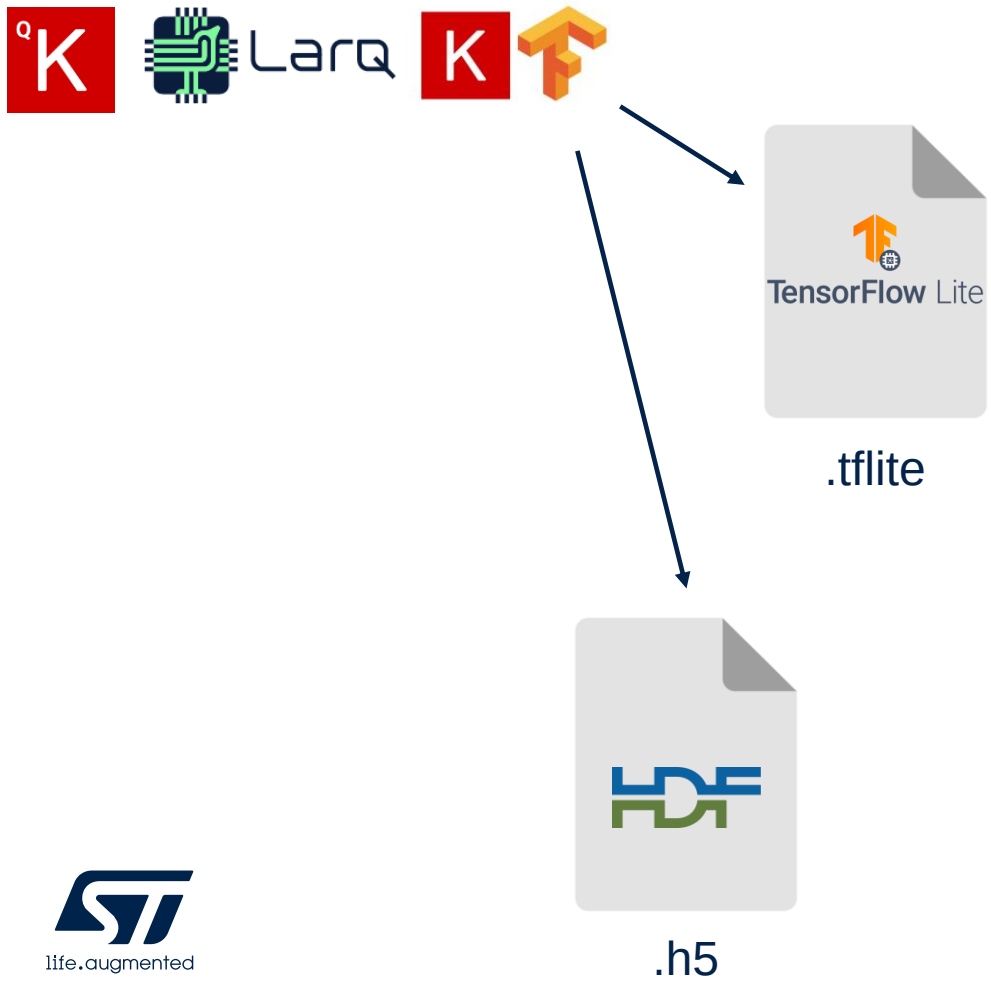
.tflite

Heterogeneity: DL Formats

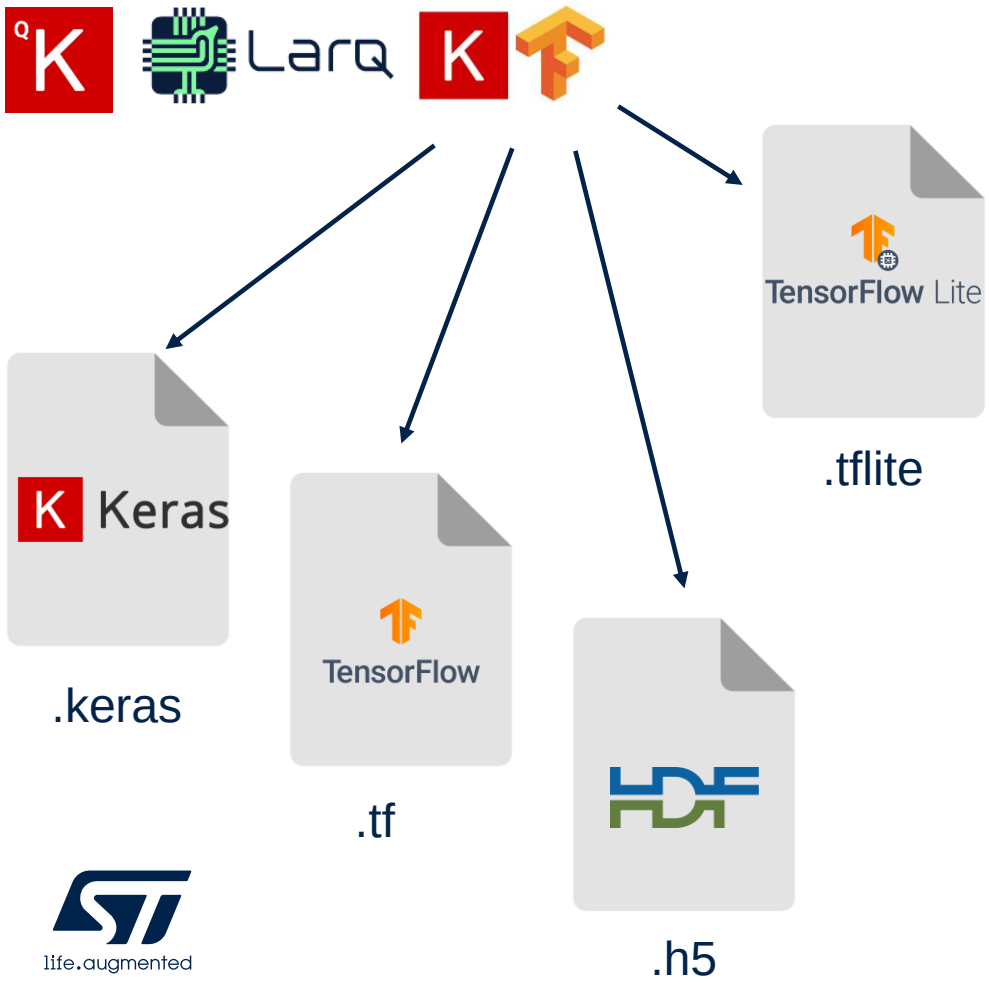


.tflite

Heterogeneity: DL Formats

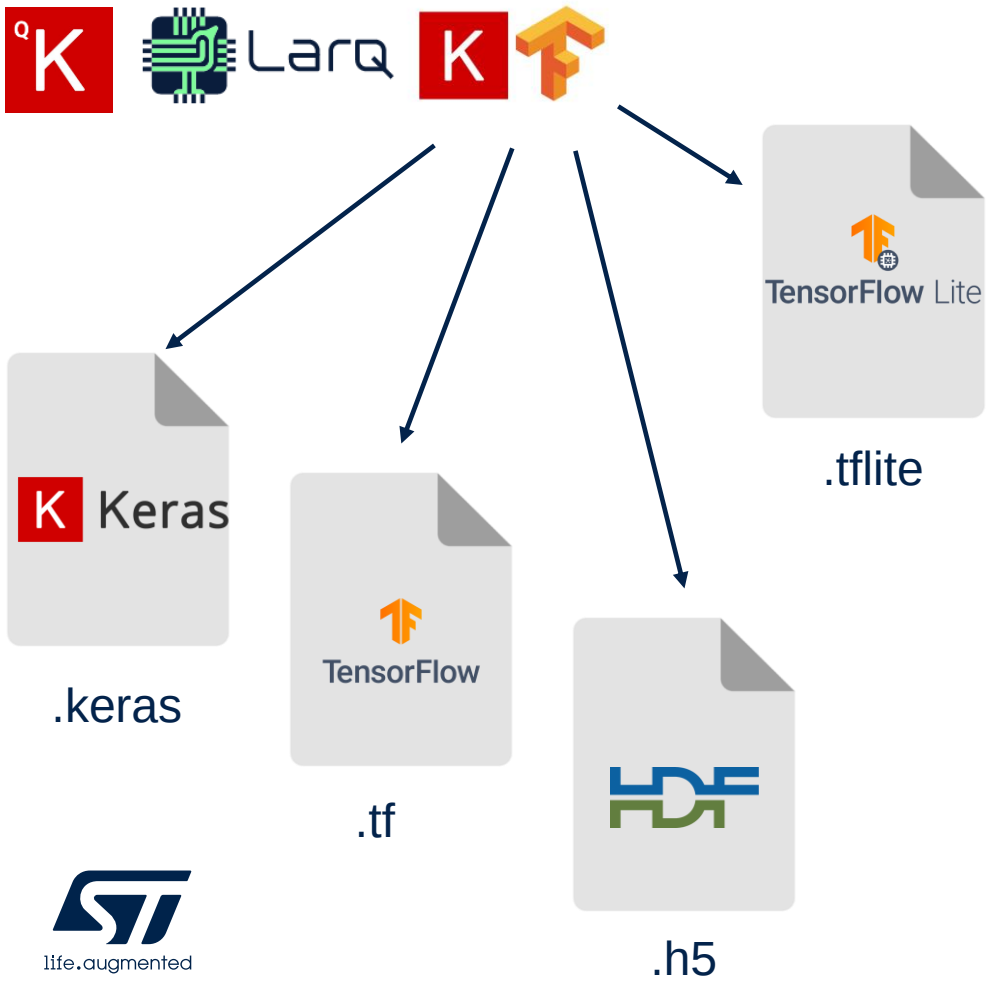


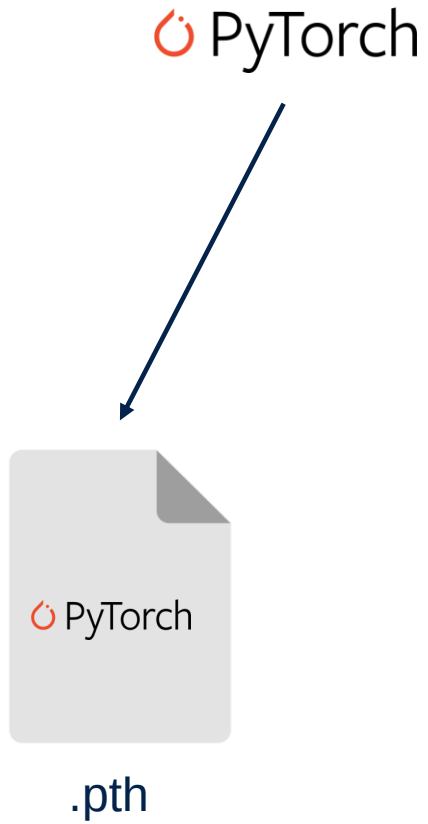
Heterogeneity: DL Formats



life.augmented

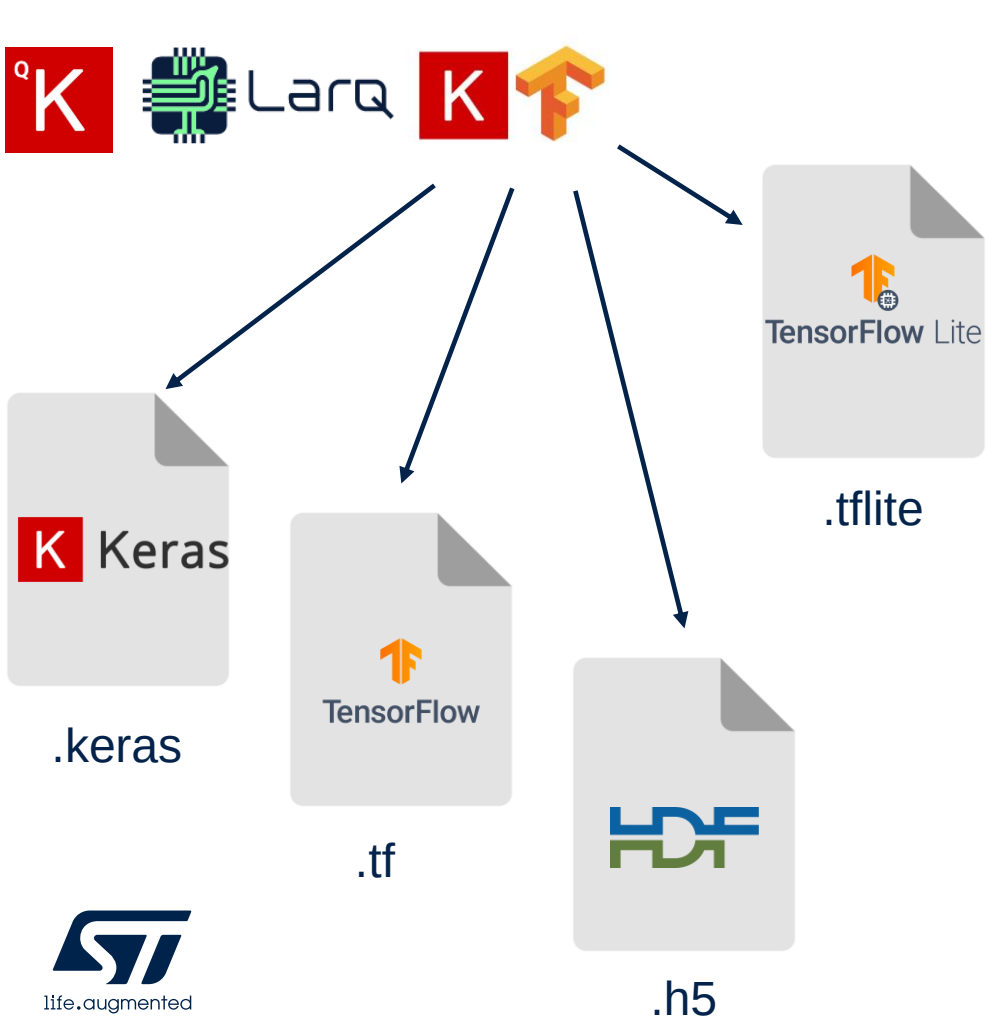
Heterogeneity: DL Formats

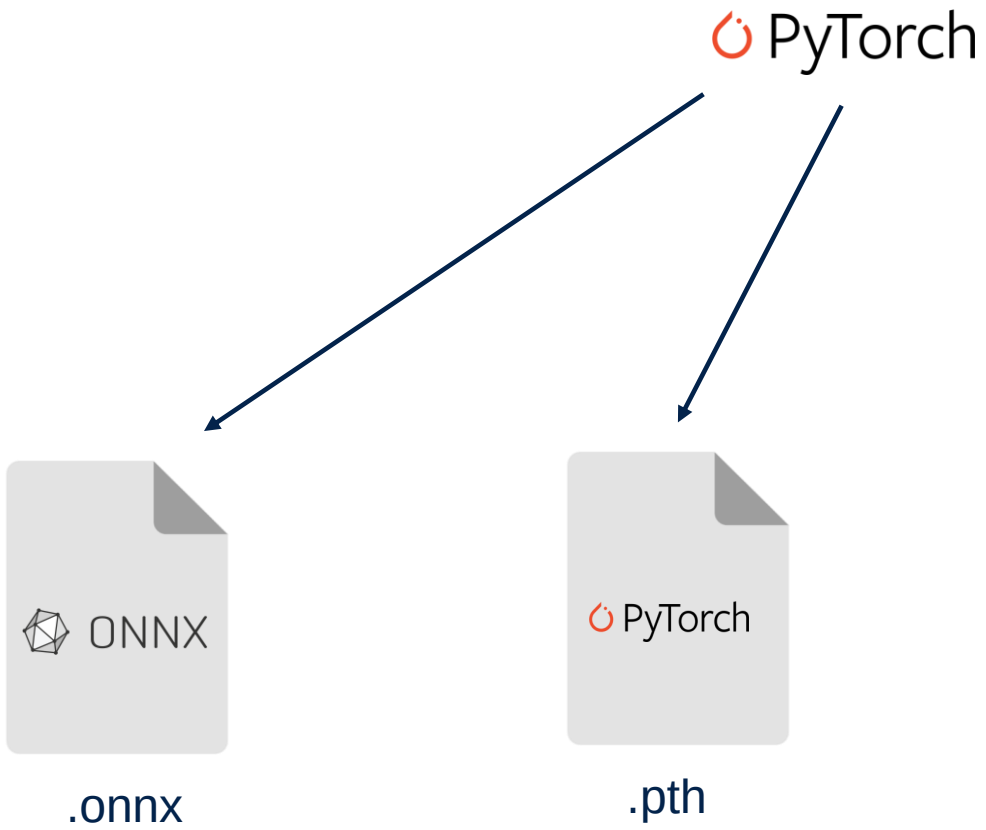




life.augmented

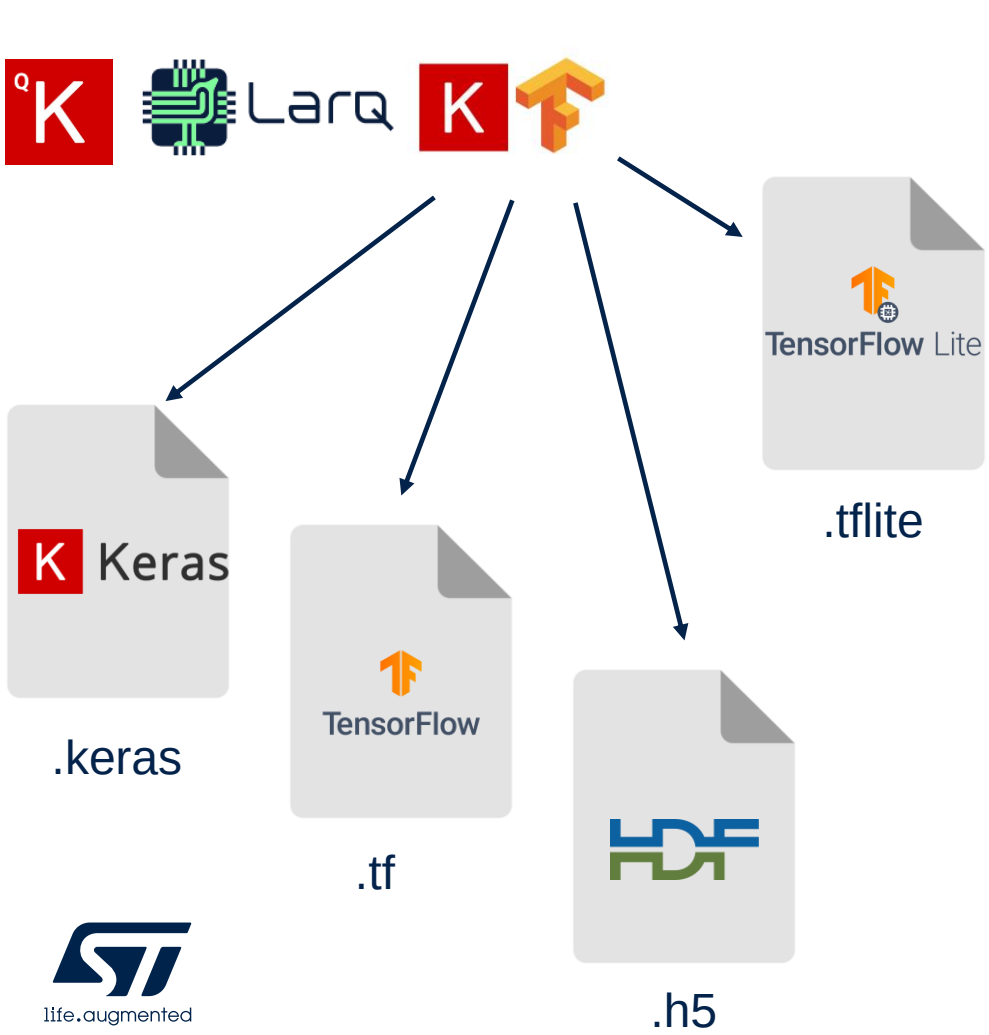
Heterogeneity: DL Formats

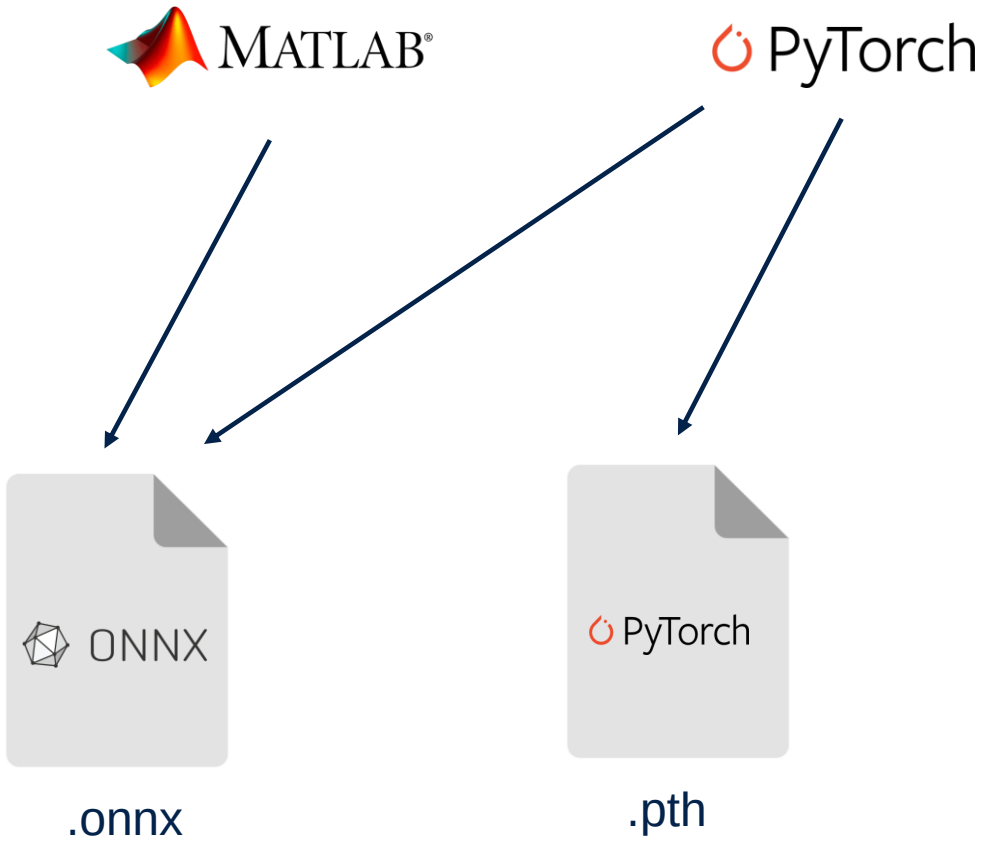




life.augmented

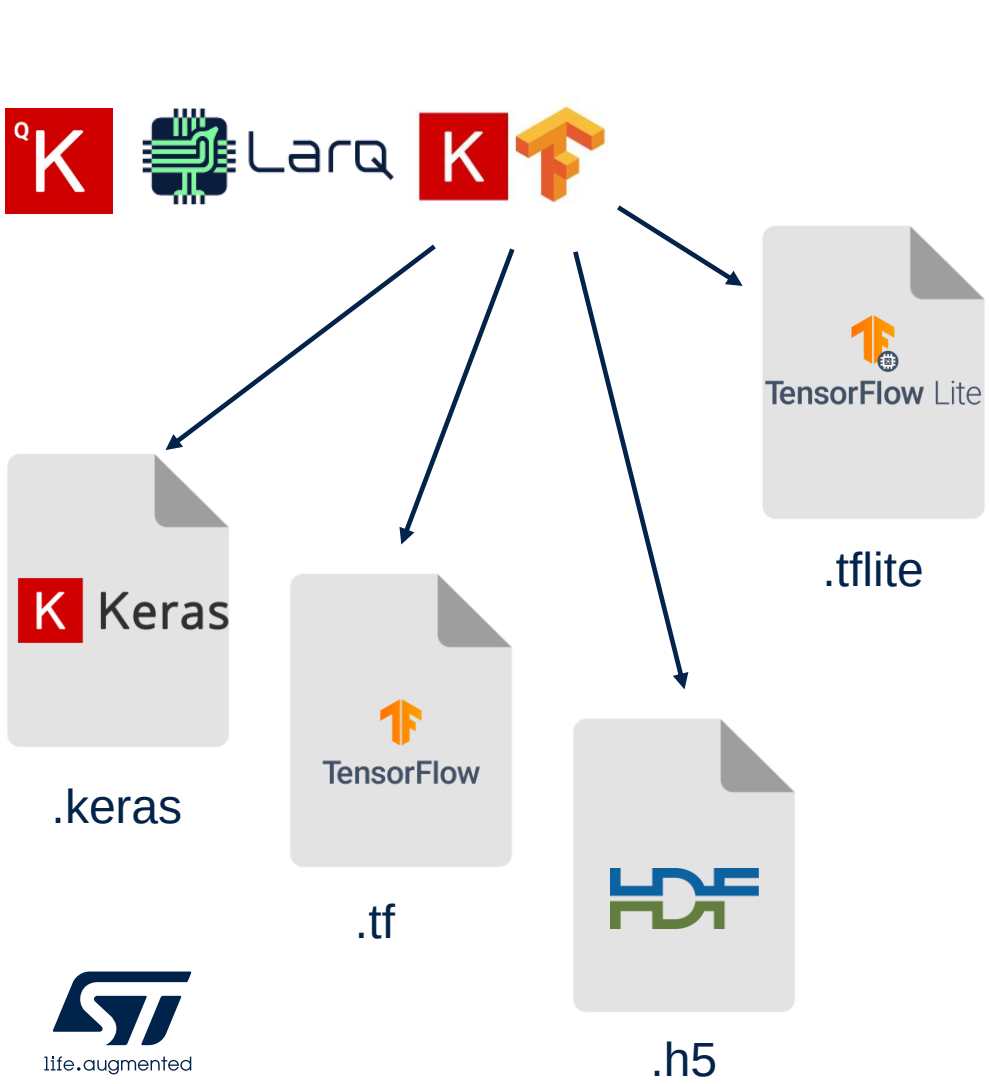
Heterogeneity: DL Formats

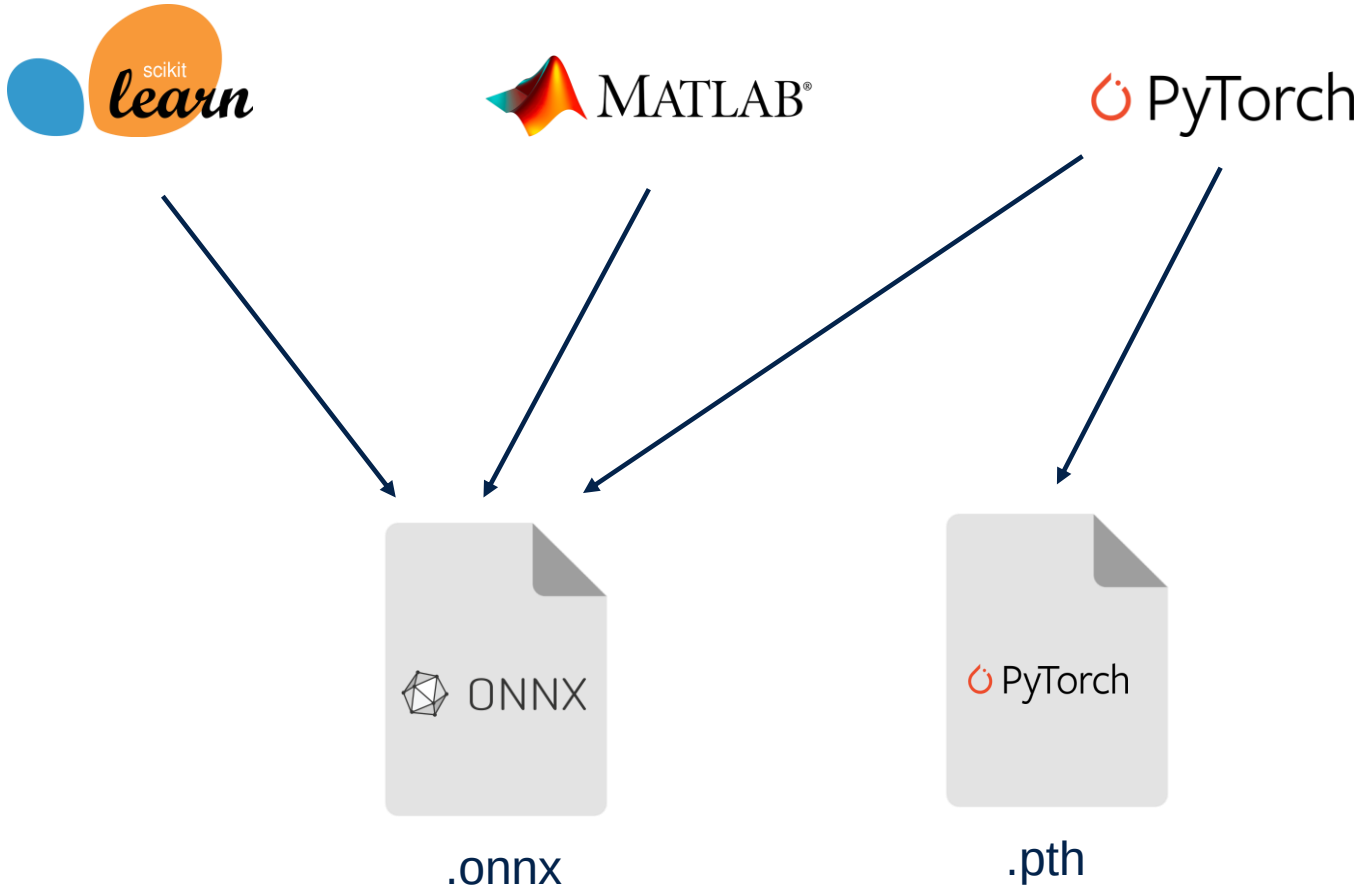




life.augmented

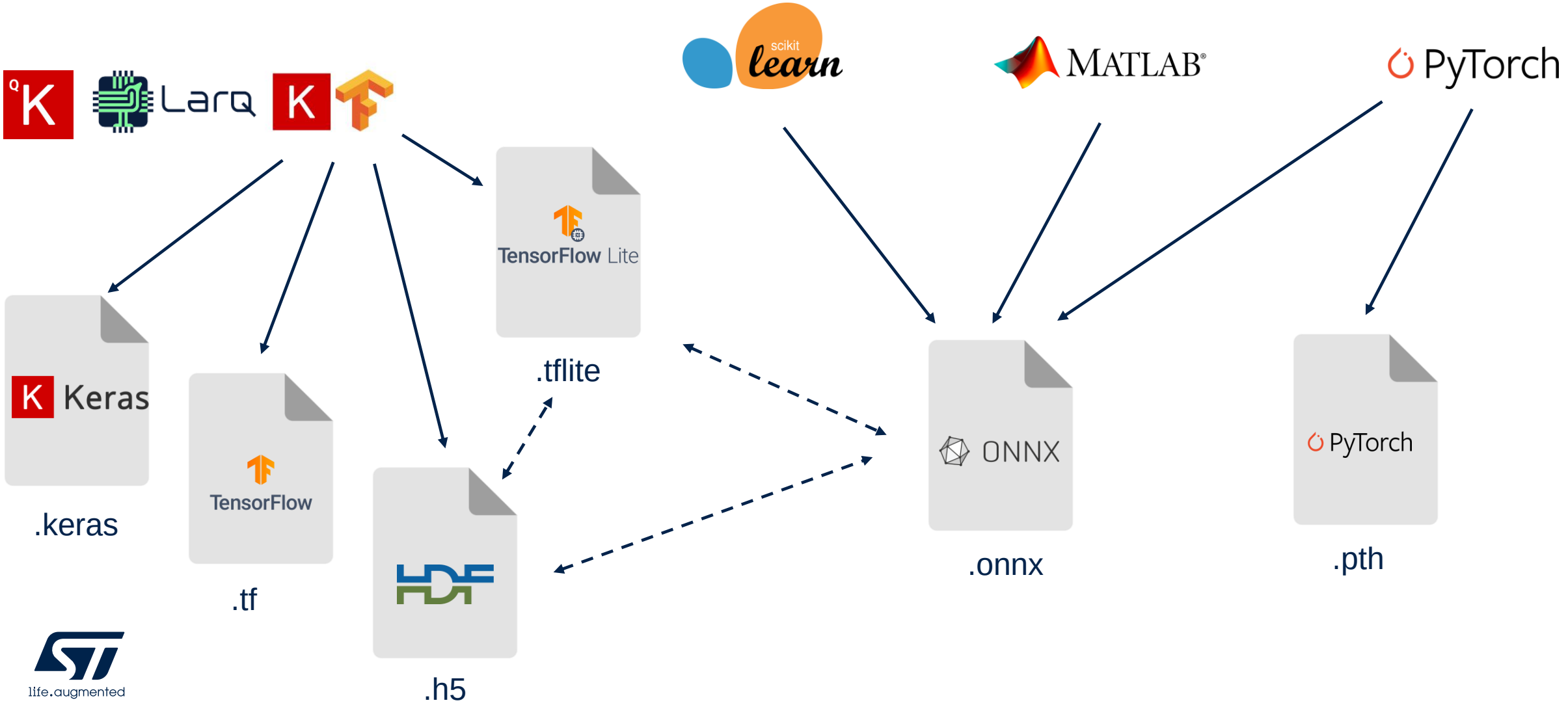
Heterogeneity: DL Formats





life.augmented

Heterogeneity: DL Formats



life.augmented

Heterogeneity: DL Differences

- Availability of operators: not all operators are available in all frameworks
- Attributes: different attributes, different meaning, different default values
- Data layout, i.e., channel first vs. channel last
- Quantization, i.e., available precisions to represent data

Heterogeneity: DL Differences

- Availability of operators: not all operators are available in all frameworks
- Attributes: different attributes, different meaning, different default values
- Data layout, i.e., channel first vs. channel last
- Quantization, i.e., available precisions to represent data

TFLite
MEAN

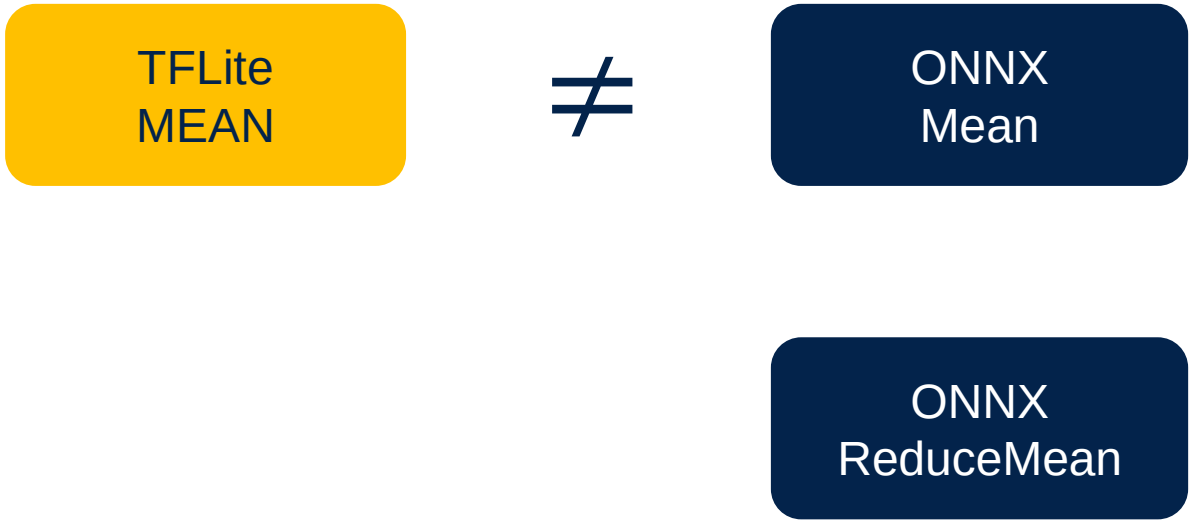
ONNX
Mean

Heterogeneity: DL Differences

- Availability of operators: not all operators are available in all frameworks
- Attributes: different attributes, different meaning, different default values
- Data layout, i.e., channel first vs. channel last
- Quantization, i.e., available precisions to represent data

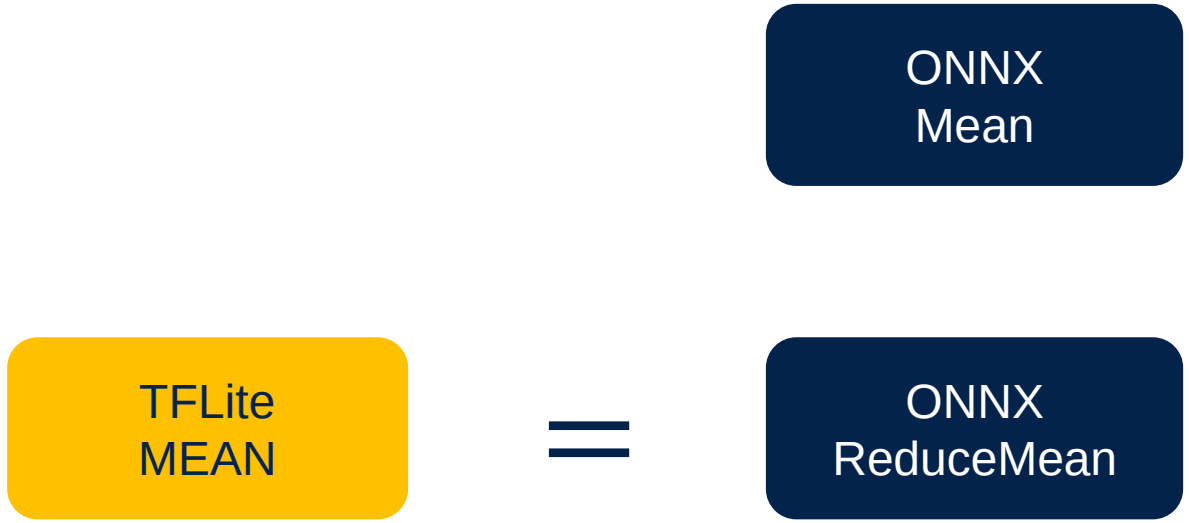
Heterogeneity: DL Differences

- Availability of operators: not all operators are available in all frameworks
- Attributes: different attributes, different meaning, different default values
- Data layout, i.e., channel first vs. channel last
- Quantization, i.e., available precisions to represent data

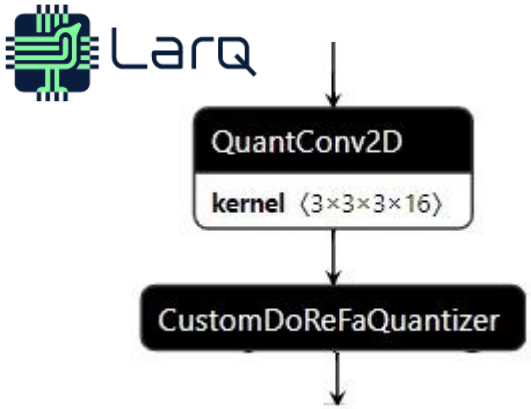


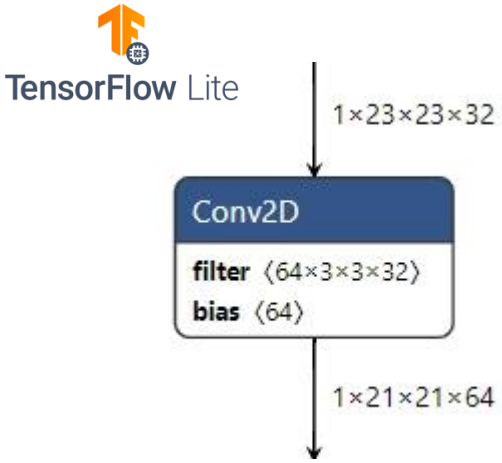
Heterogeneity: DL Differences

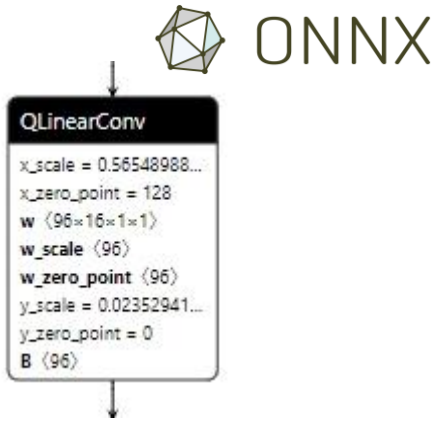
- Availability of operators: not all operators are available in all frameworks
- Attributes: different attributes, different meaning, different default values
- Data layout, i.e., channel first vs. channel last
- Quantization, i.e., available precisions to represent data

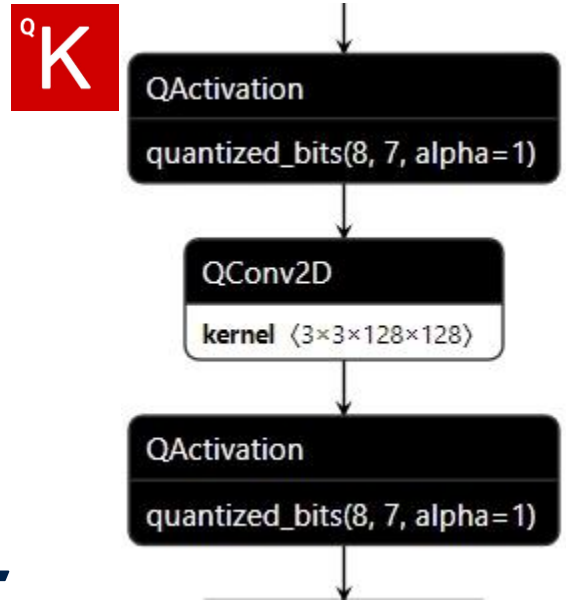


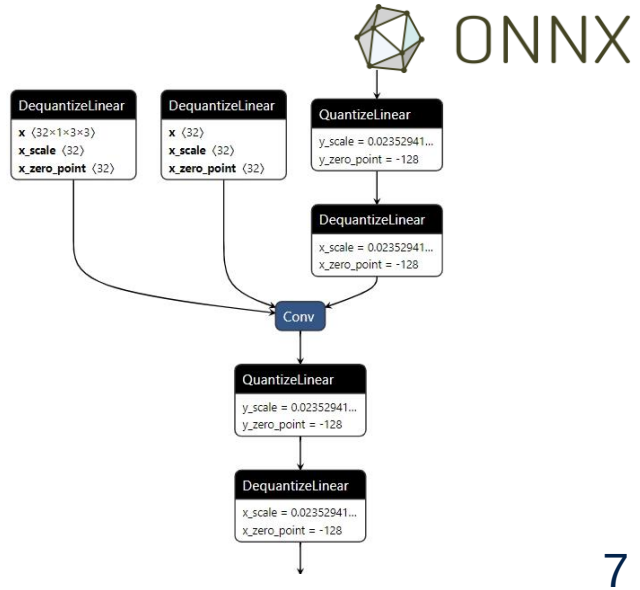
Heterogeneity: DL Formats





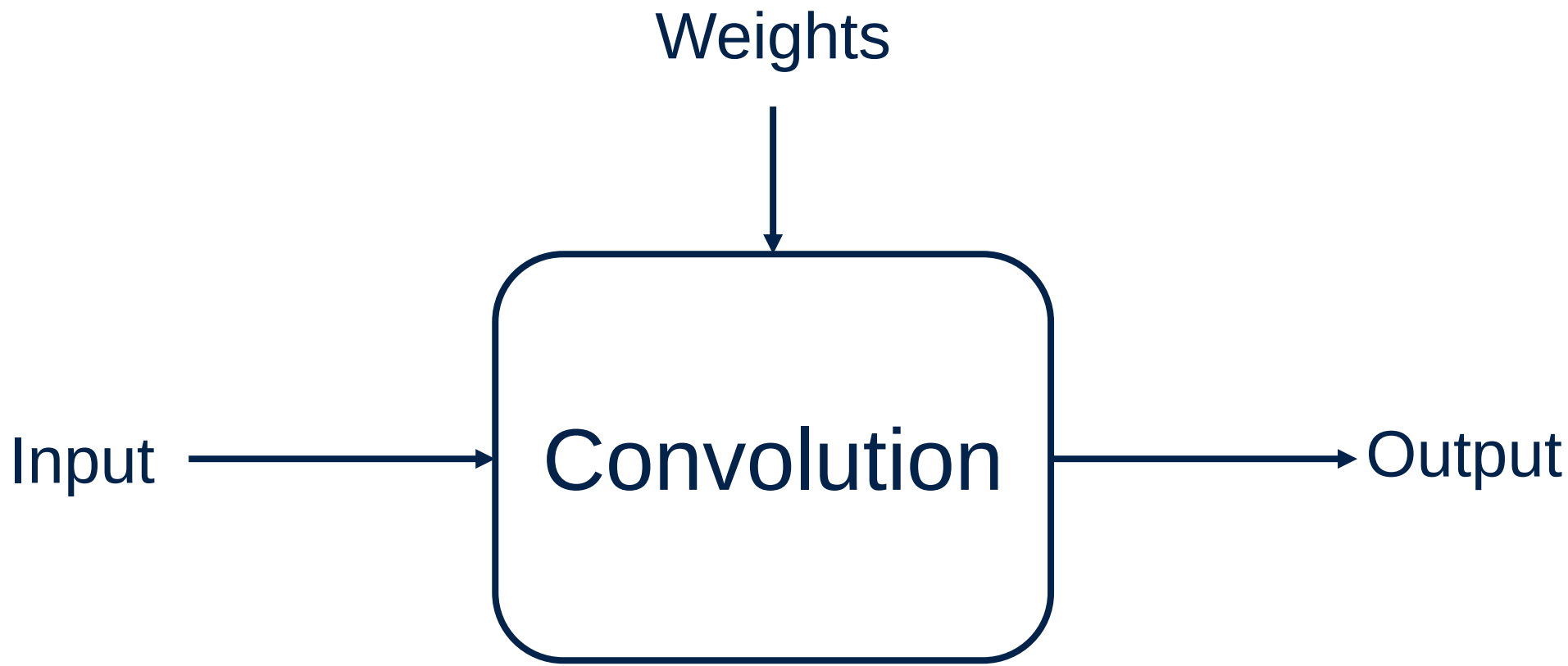




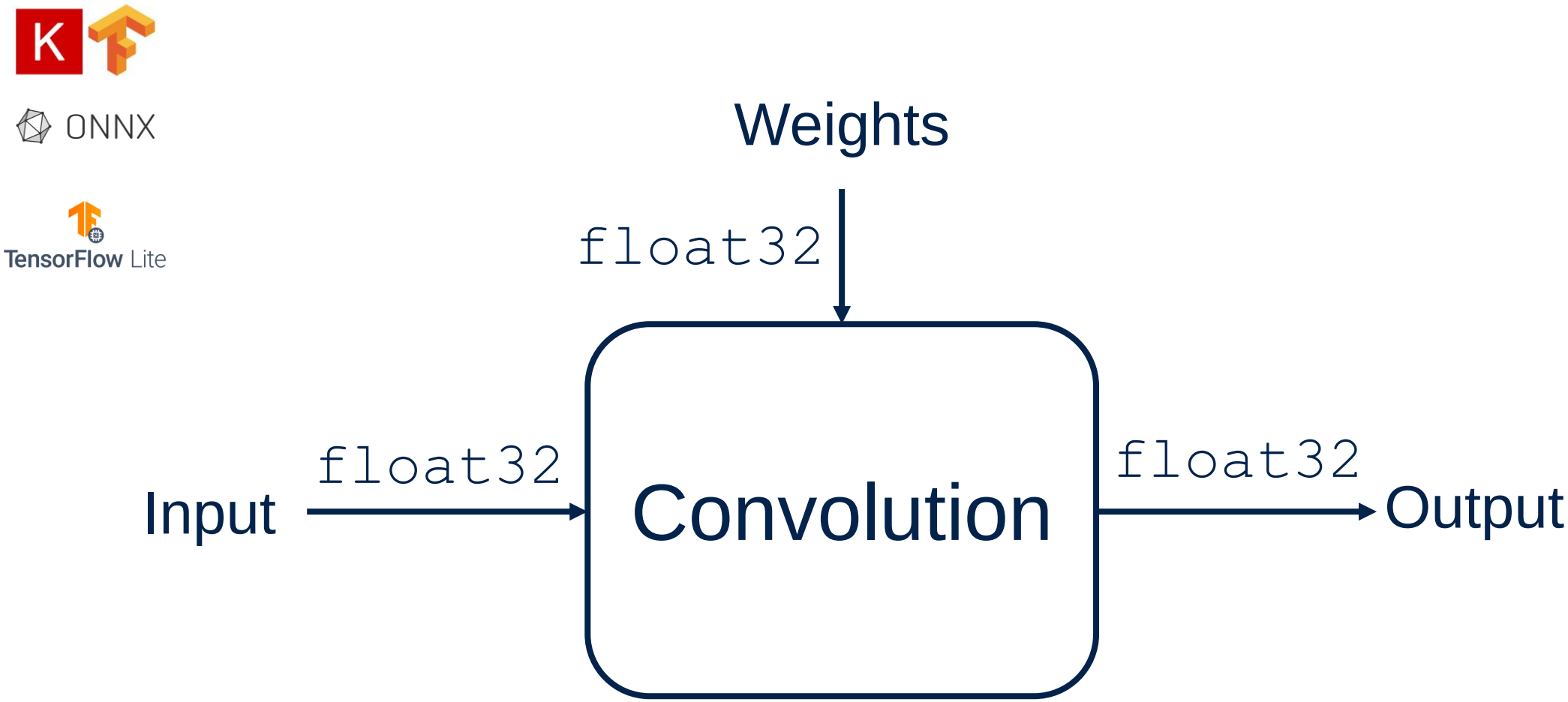


life.augmented

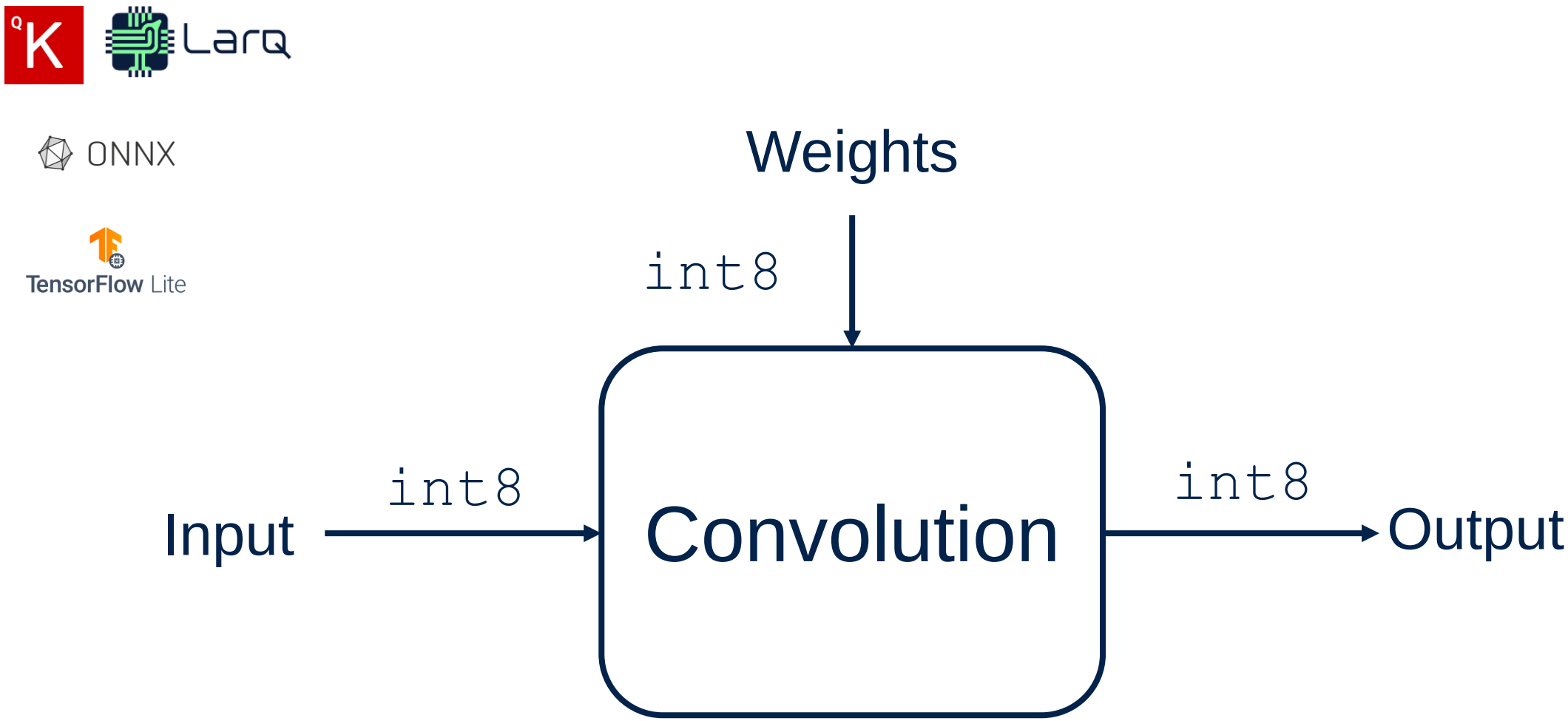
Heterogeneity: Quantization



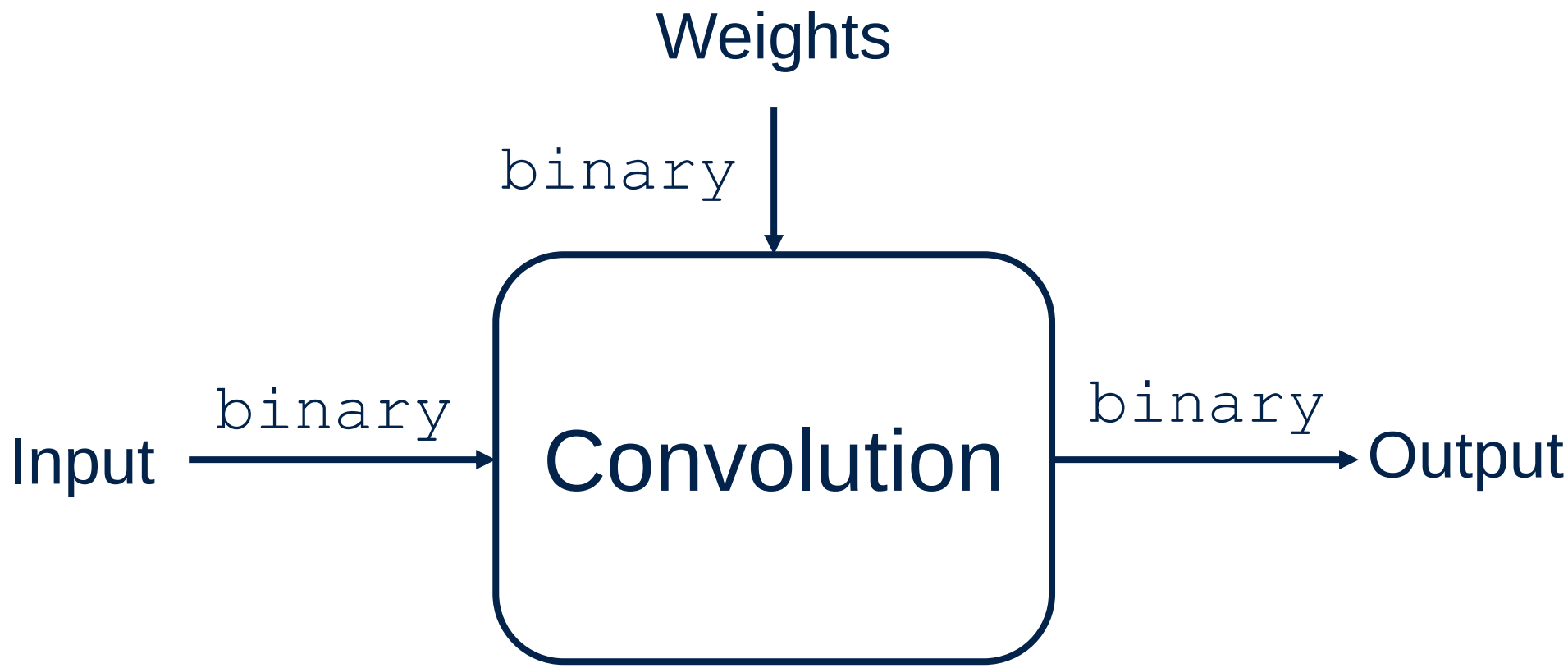
Heterogeneity: Quantization



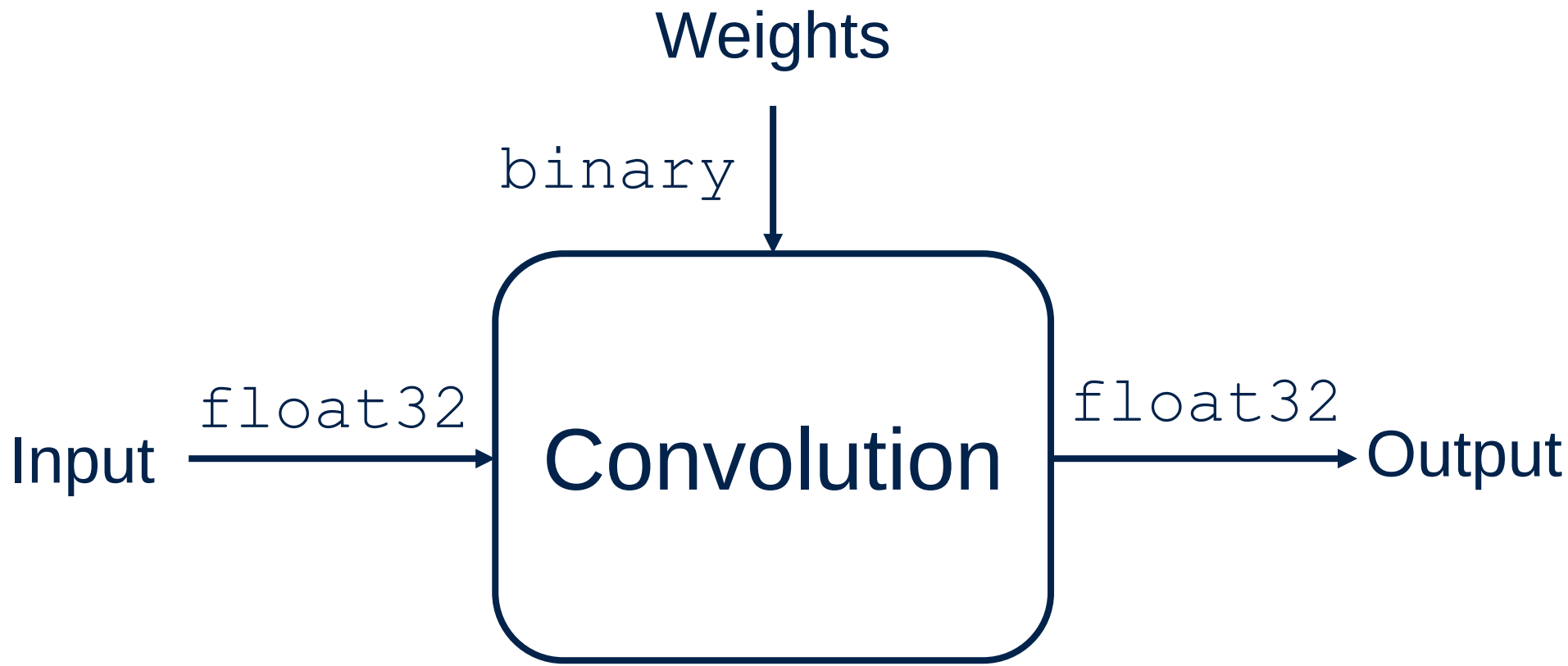
Heterogeneity: Quantization



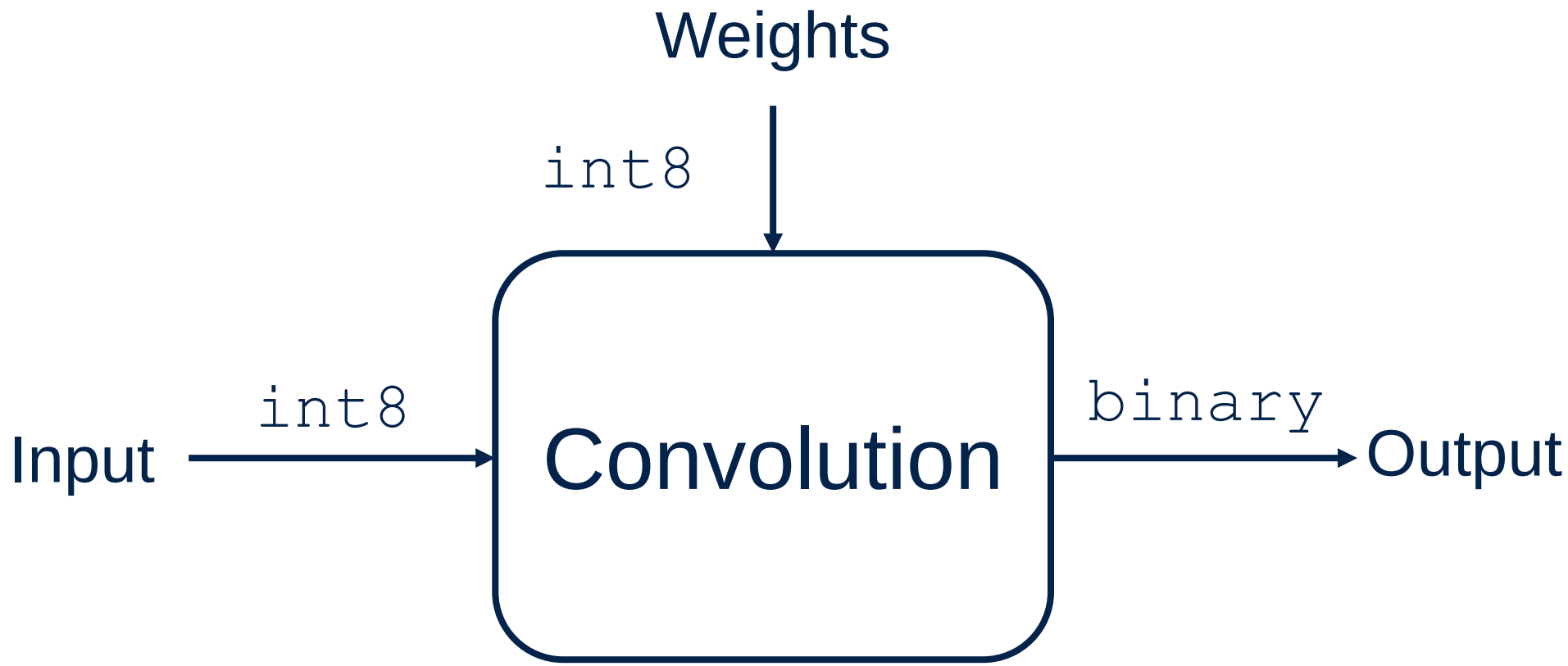
Heterogeneity: Quantization



Heterogeneity: Quantization



Heterogeneity: Quantization



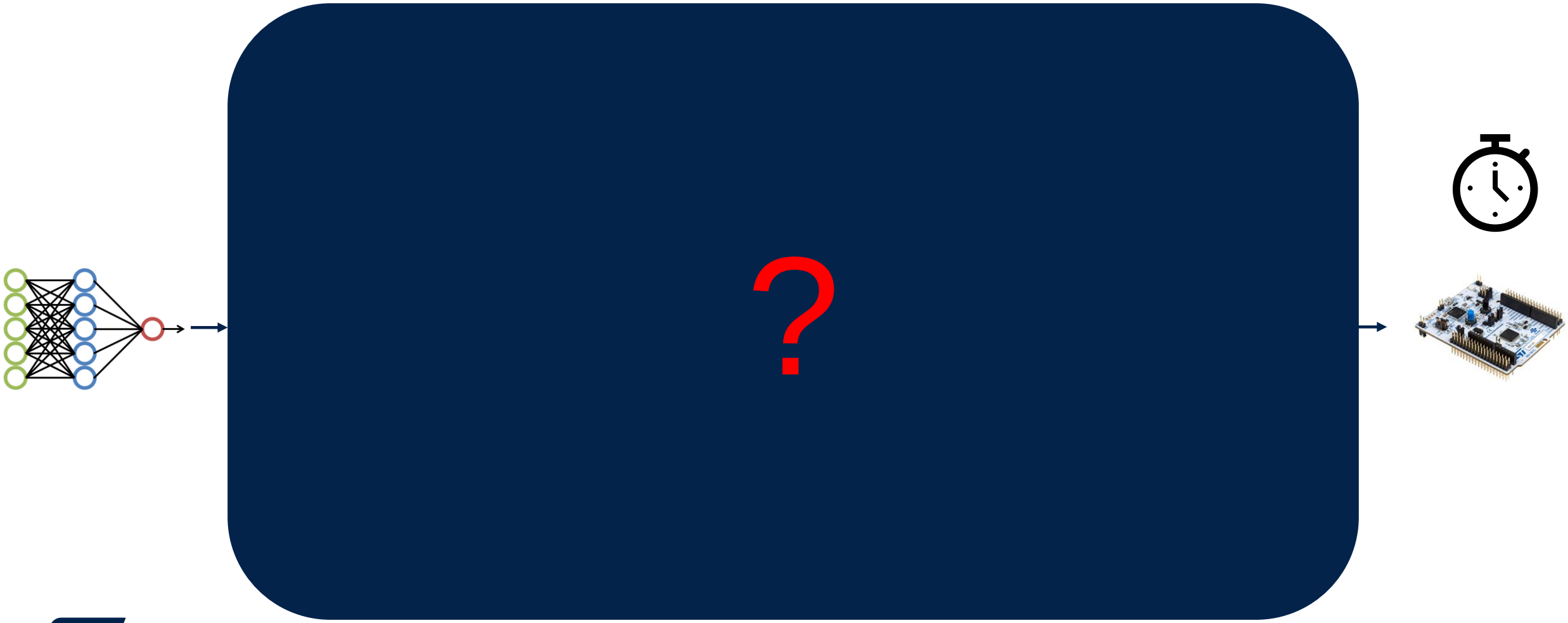
Heterogeneity: Quantization

Limiting to binary, int8, float32 there are 3^3=27 combinations to be supported for a single layer

Input

put

Removing Heterogeneity



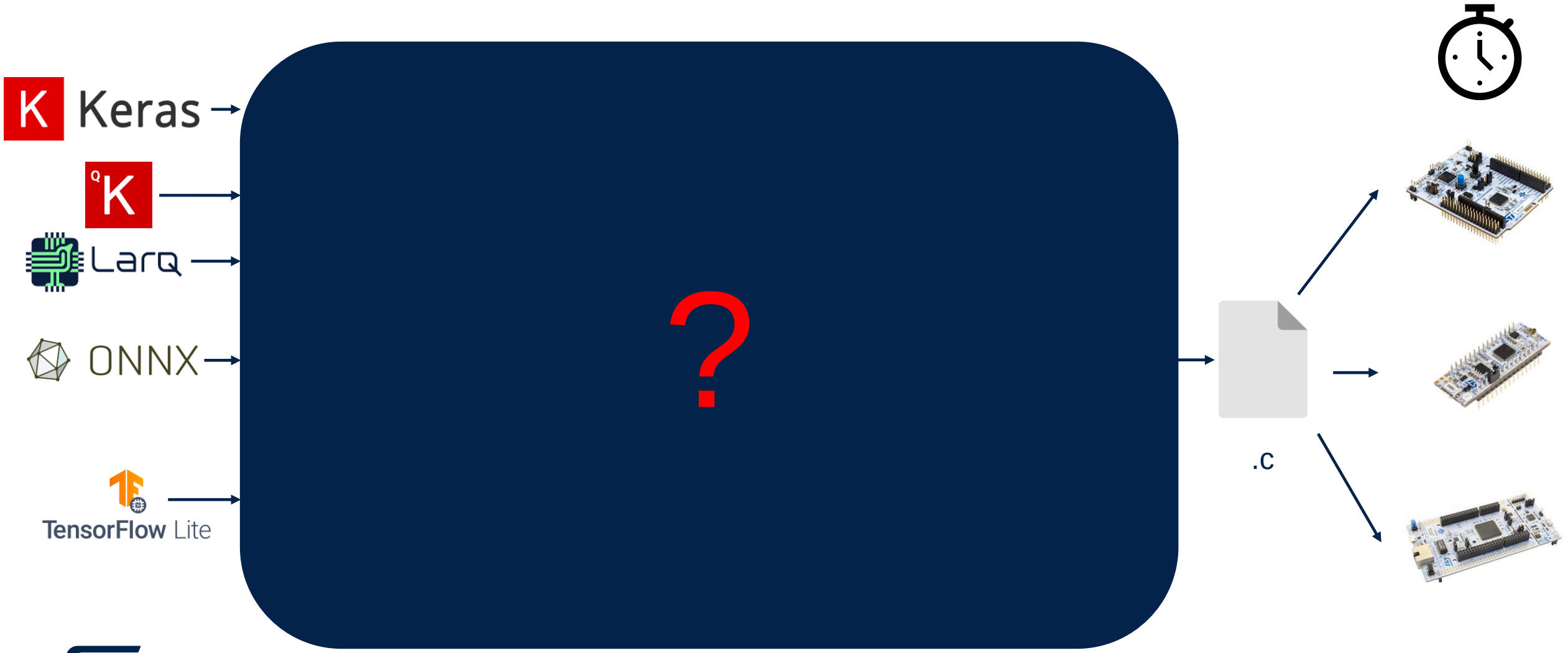
life.augmented

Removing Heterogeneity

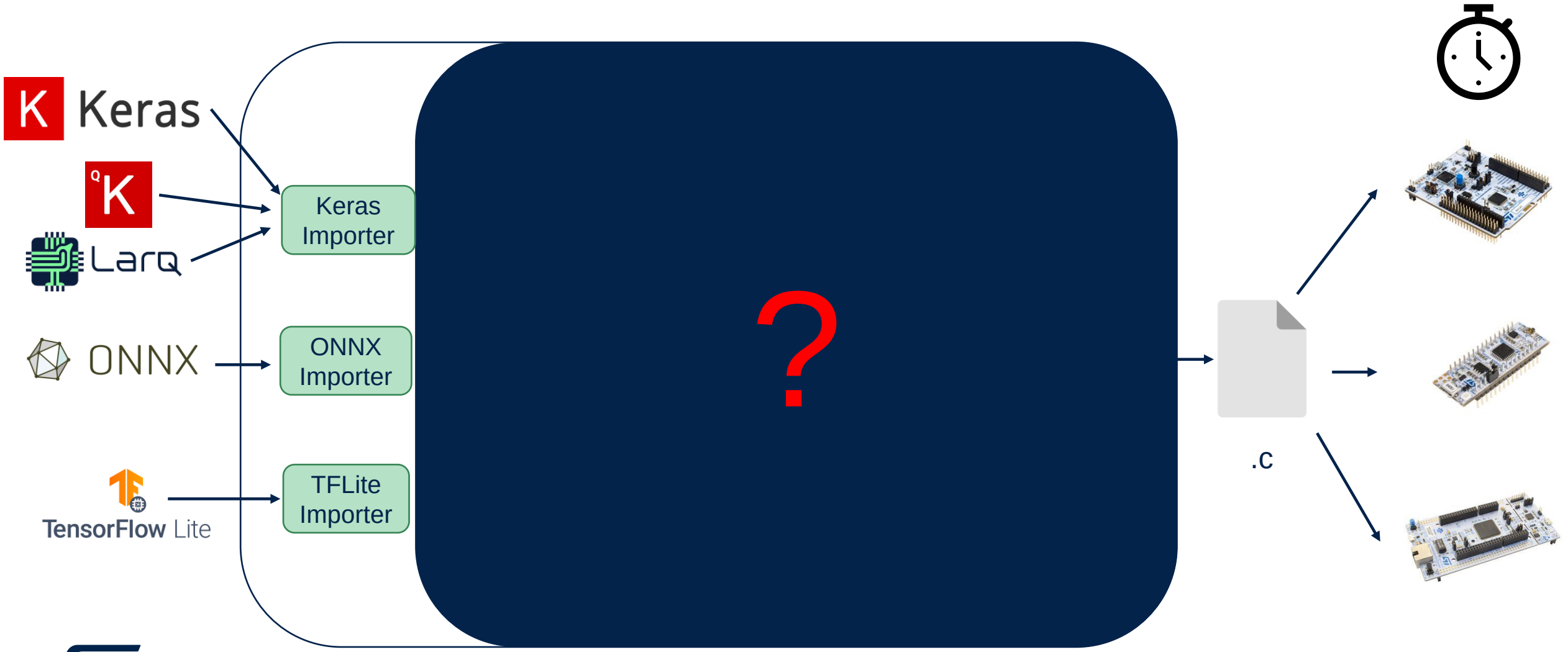
Removing Heterogeneity

life.augmented

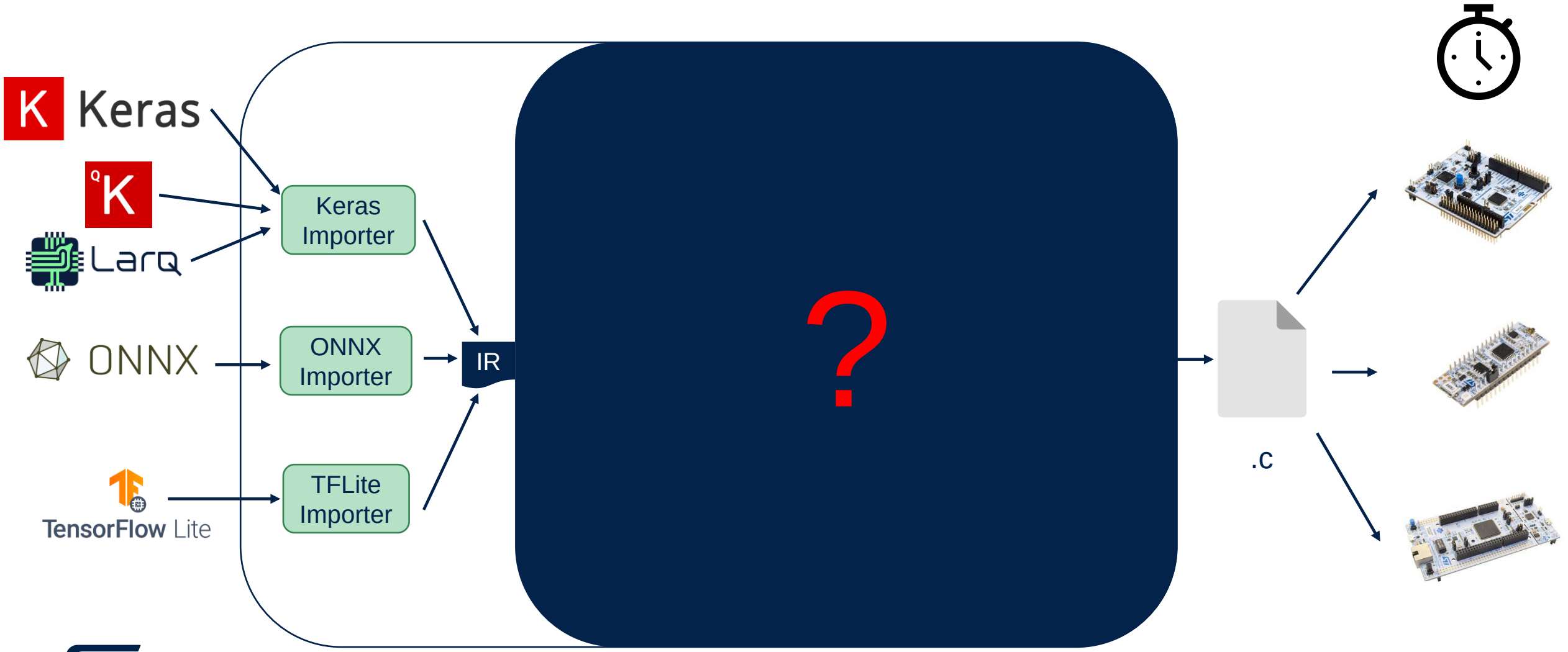
Removing Heterogeneity



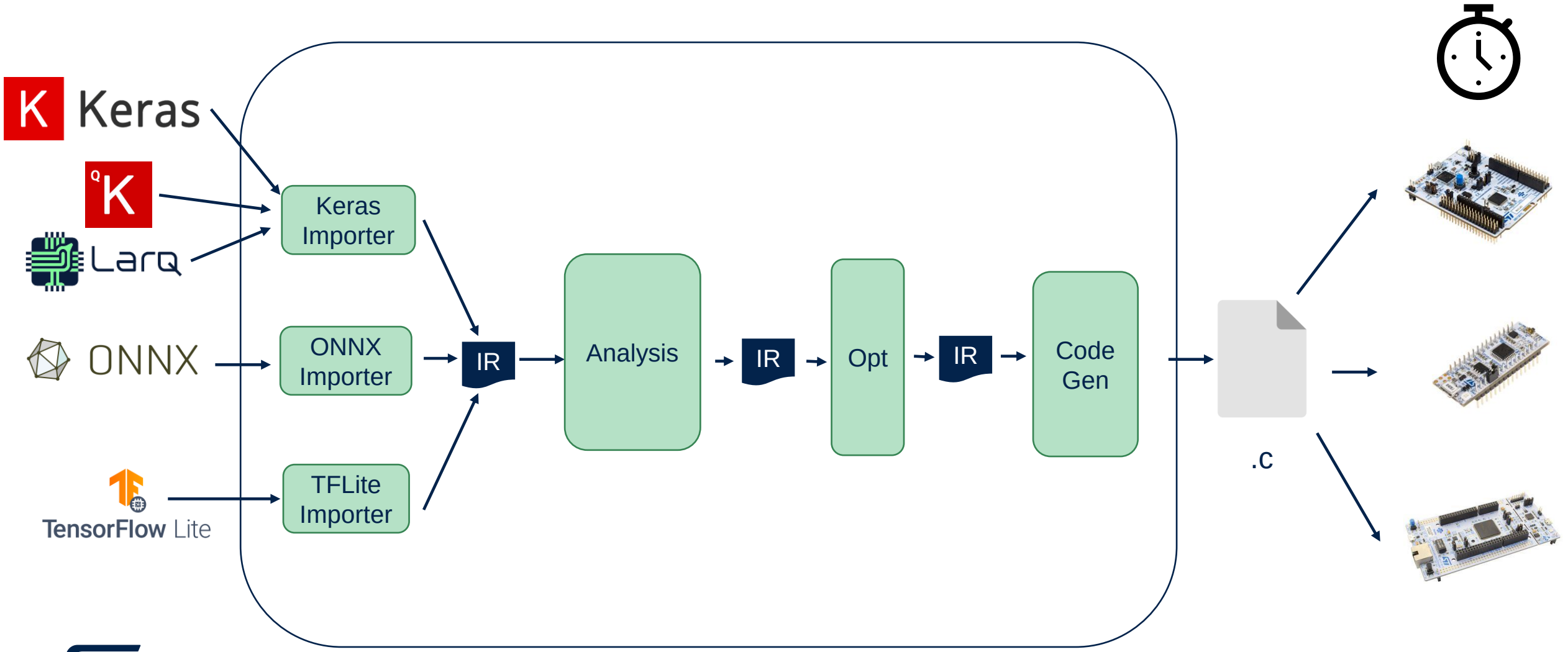
Removing Heterogeneity



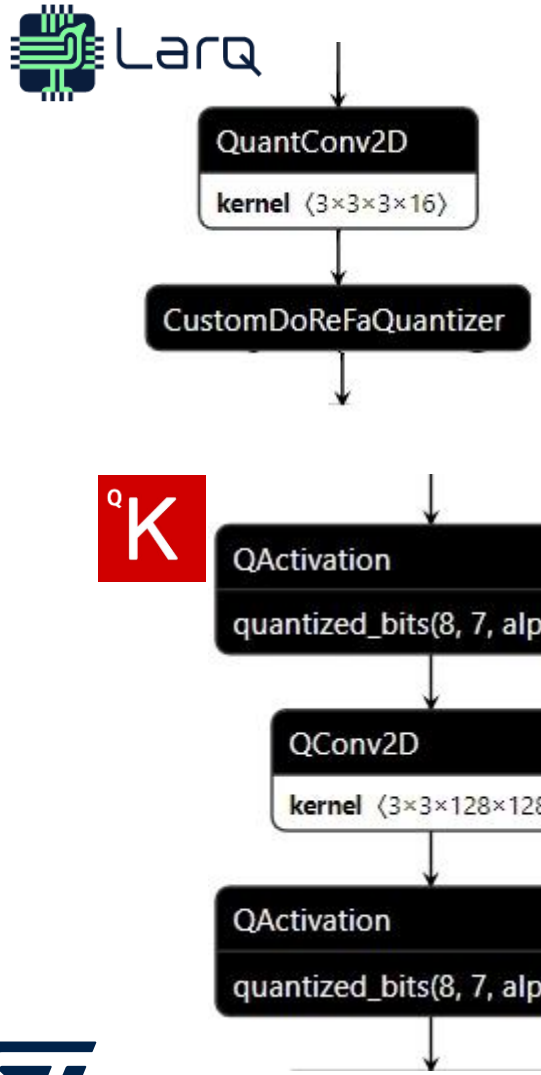
Removing Heterogeneity

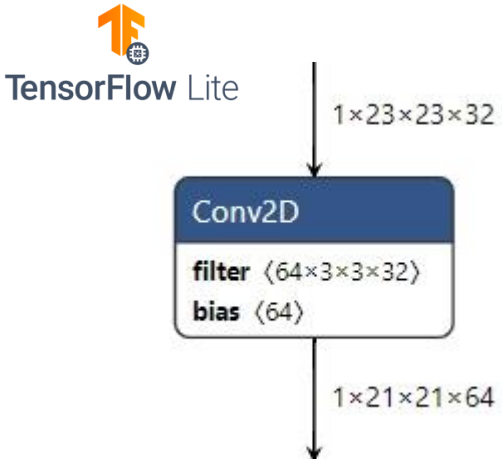


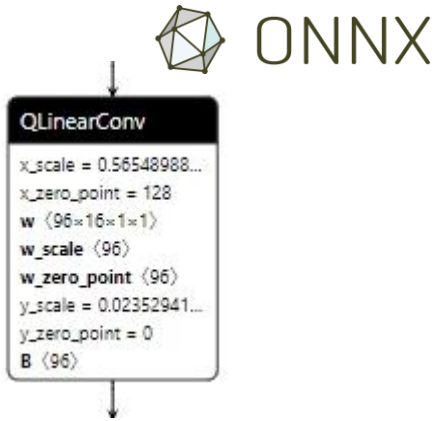
Removing Heterogeneity

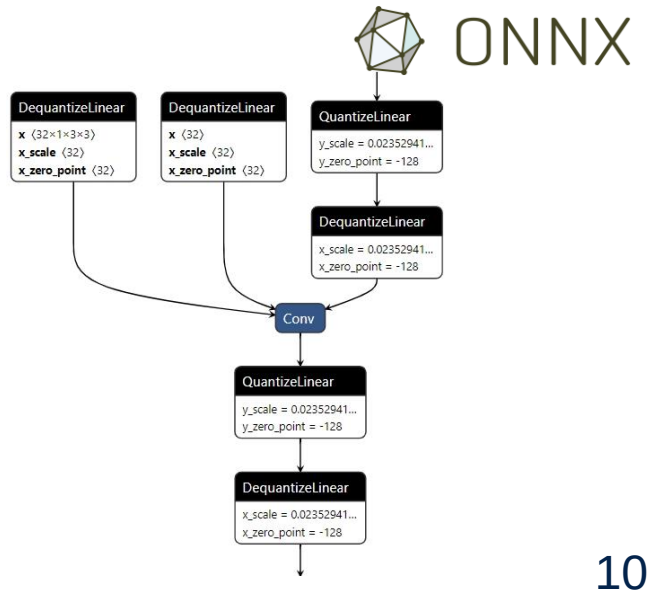


Common Intermediate Representation



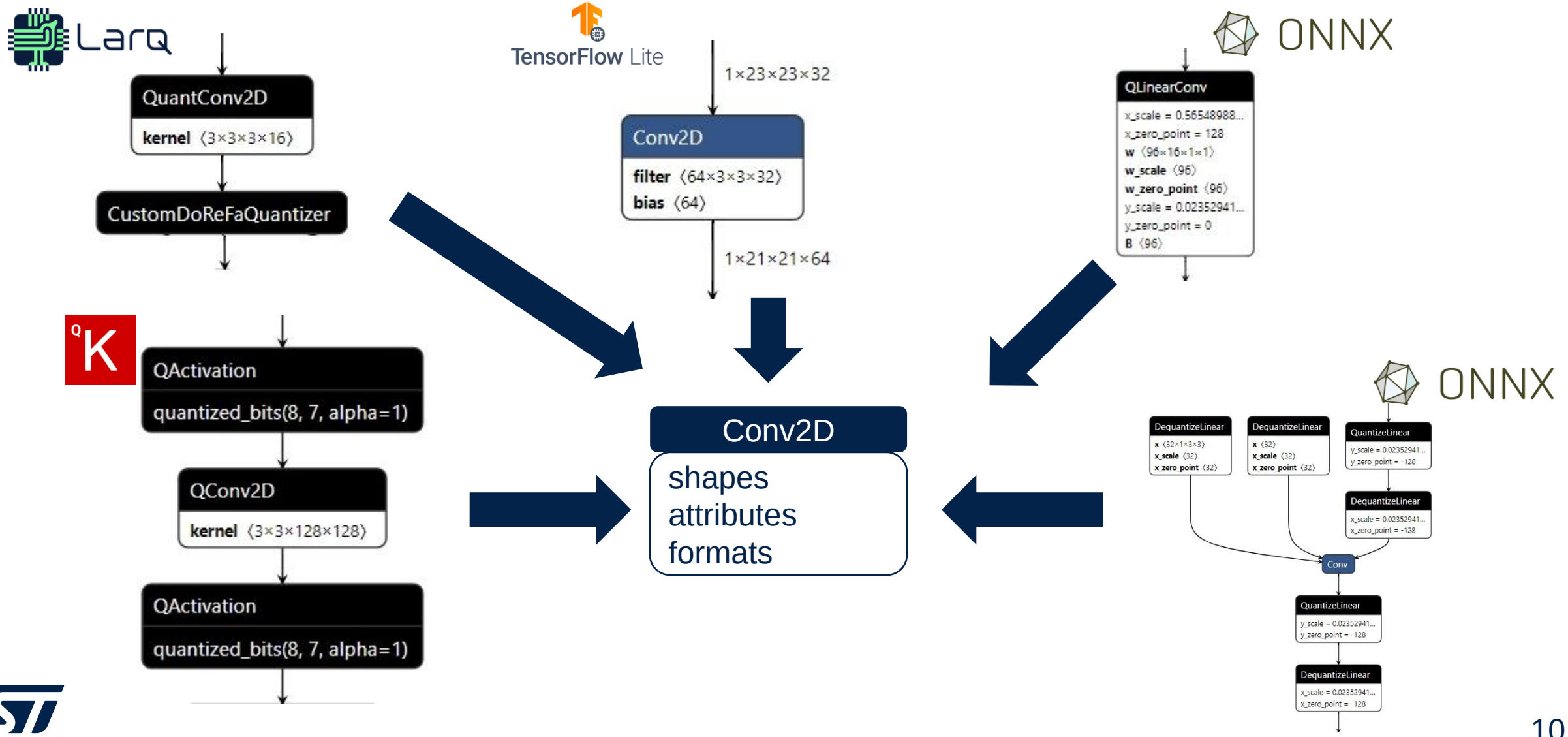




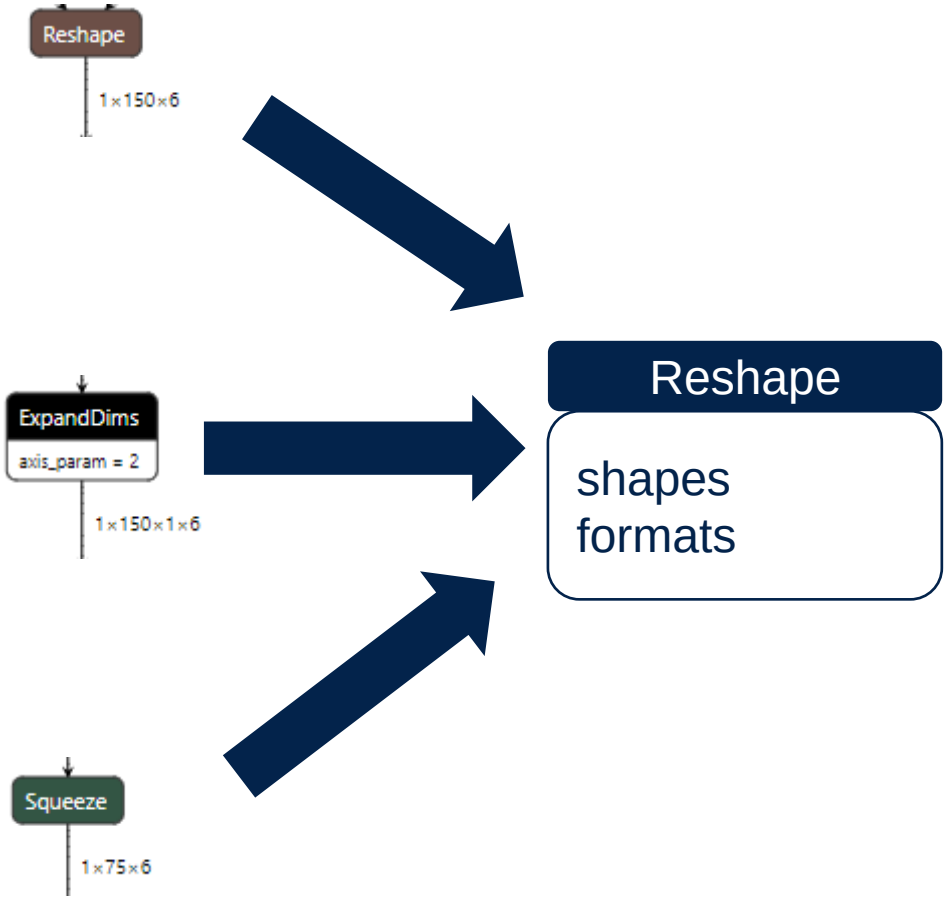


life.augmented

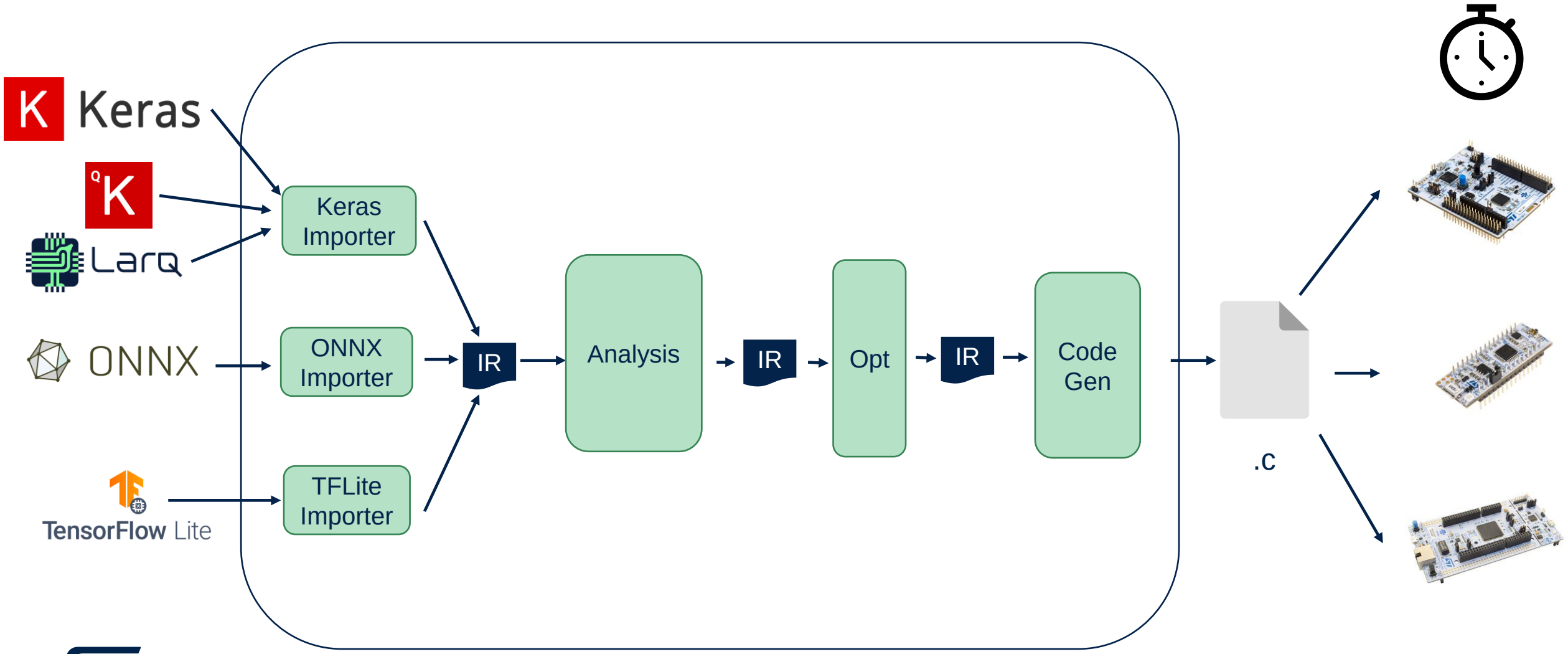
Common Intermediate Representation



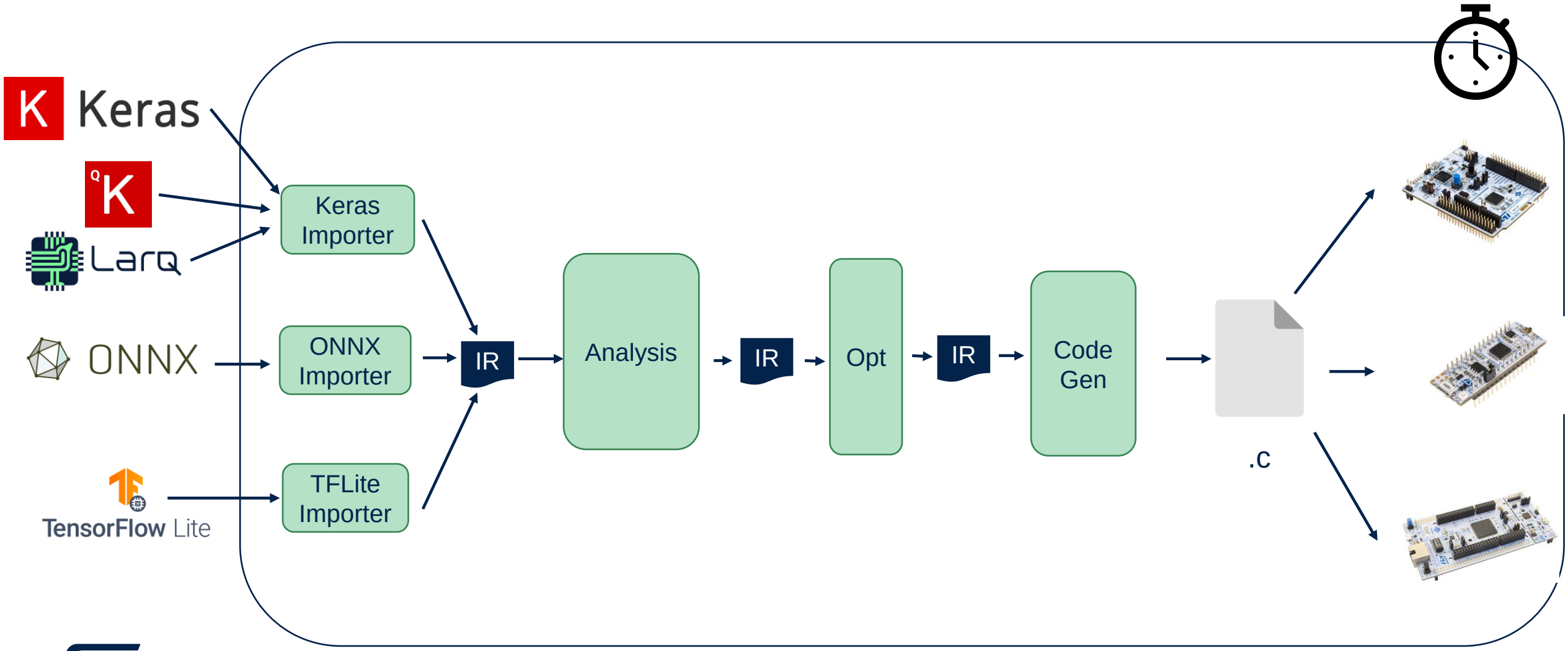
Common Intermediate Representation



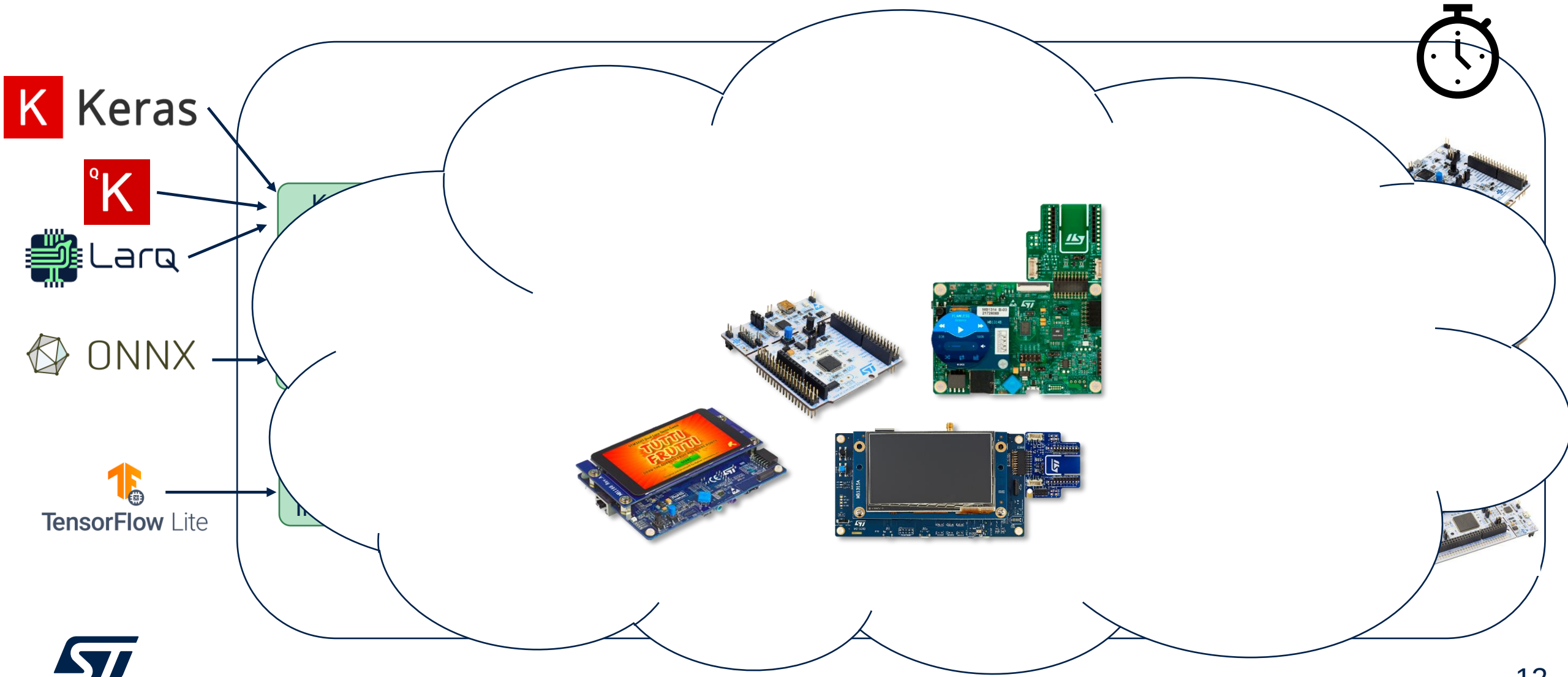
STM32Cube.AI Developer Cloud



STM32Cube.AI Developer Cloud

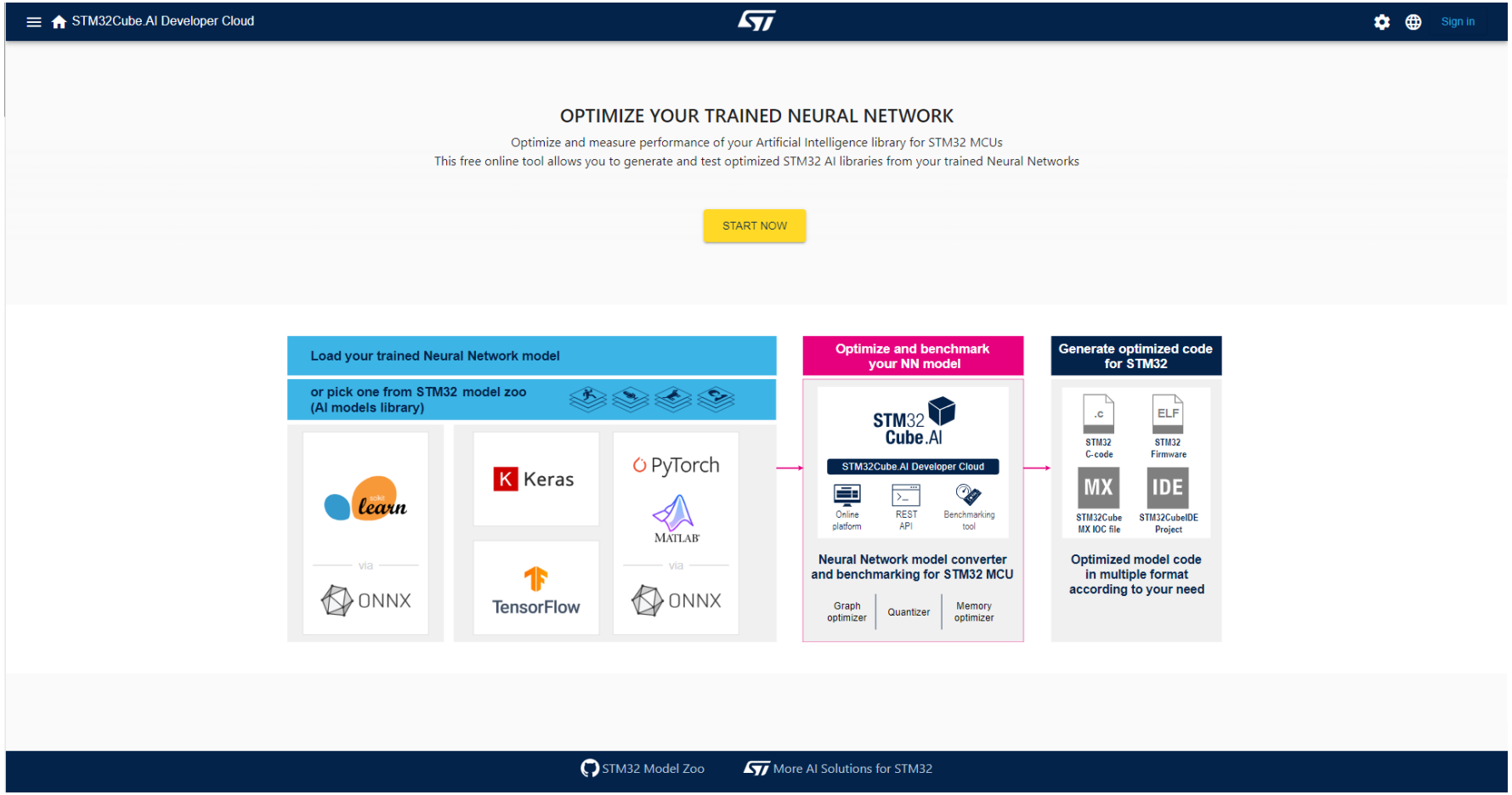


STM32Cube.AI Developer Cloud



life.augmented

STM32Cube.AI Developer Cloud



https://stm32ai-cs.st.com/home

life.augmented

Our technology starts with You

Find out more at www.st.com

life.augmented