aidge

The first independent, open-source platform dedicated to embedded AI

FRANCE 2030

EU

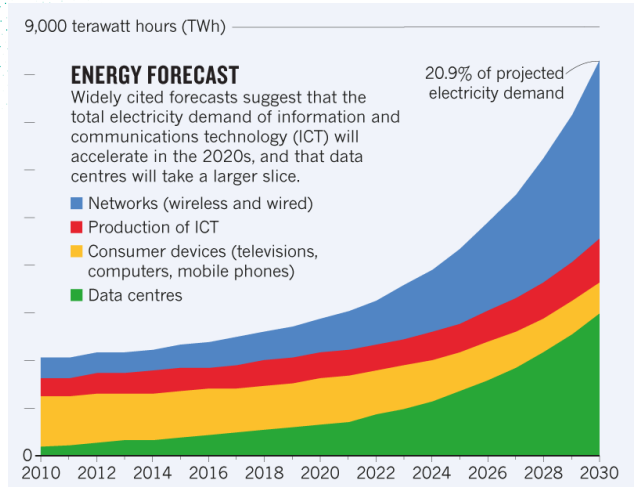
Embedded AI : deep digitial transformation

OPPORTUNITY

- Real time
- Data and Model Security
- Cost reduction

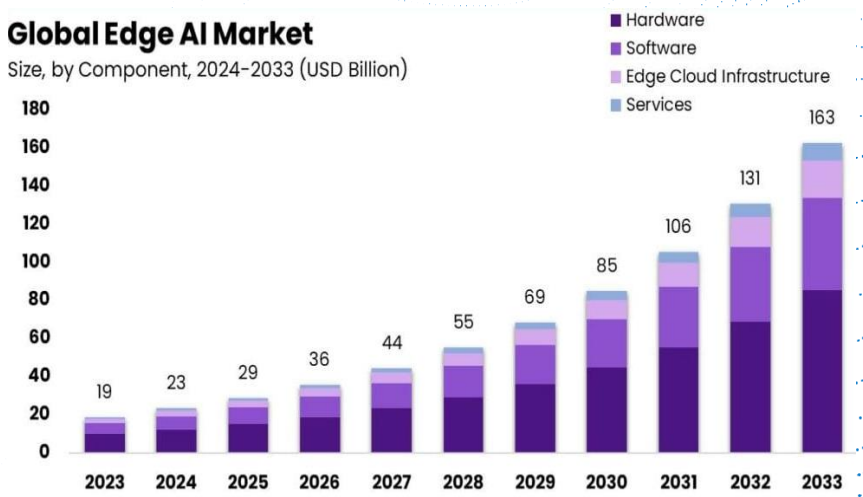
NECESSITY

- Scaling up AI
- Economic Challenge
- Strategic challenge



Communication | Data center

[Nature]



[Market.us]

Existing tools : strong orientations

Deep Learning Plateforms

TensorFlow Lite

K

• WARNING

Quantization is in beta and subject to change.

Close source / Maturity
Dependencies

Compilers

tvm

OpenXLA

Low level optimization
Black box

Hardware SDK

TensorRT

Qualcomm

ONNX RUNTIME

Close source
Hardware spectific

Challenges

OPENESS

Reusable and adaptable tools to foster innovation

INDEPENDENCE

Free choice of material from component to system

INNOVATION

Integrating innovative paradigms for trust and frugality

Complementary consortium

aidge

IA

cea
Inria
ADAGOS
cortAIx
ezako time series solutions

Embedded systems

sysnav
THALES Building a future we can all trust
MBDA
cea
ONERA

aidge

End users

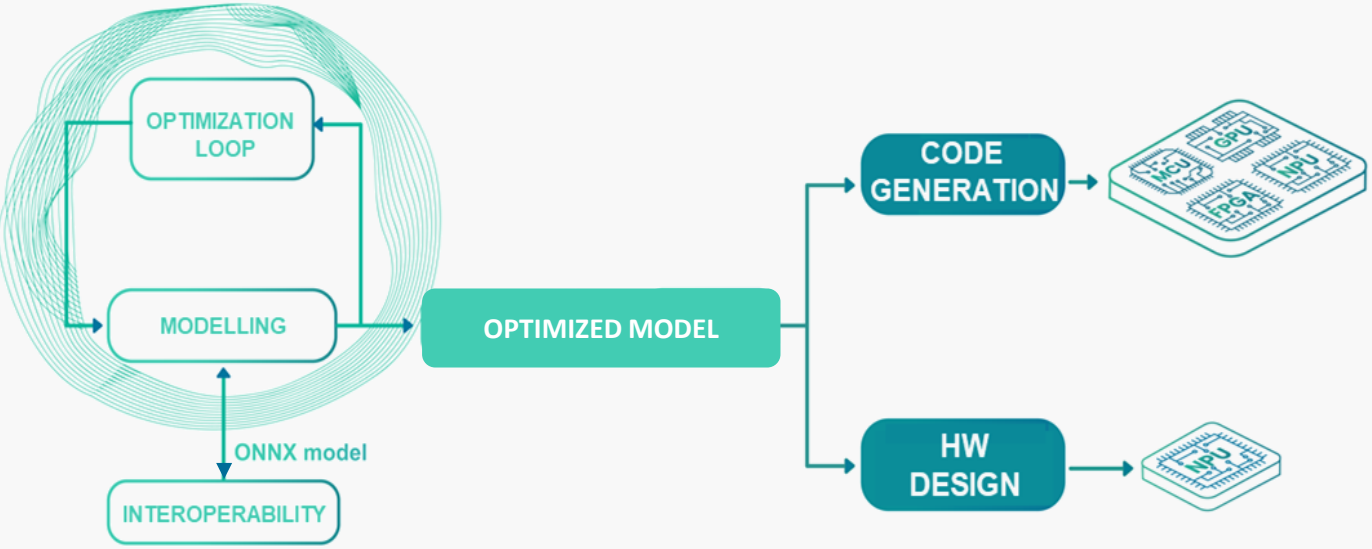
THALES Building a future we can all trust
ArcelorMittal
CS a Sopra Steria company
MBDA
PULSE AUDITION
DASSAULT AVIATION
AIRBUS
EDF

Hardware

ARCYS
HAWAI.tech HARDWARE FOR AI
KALRAY

aidge

The first open, independent platform dedicated to embedded AI



ONNX

PyTorch TensorFlow K

Integrated platform

- From import to deployment
- High degree of interoperability
- Minimal dependencies

Modular and expandable platform

- Lightweight core module with plug-ins
- Open-source collaborative environment

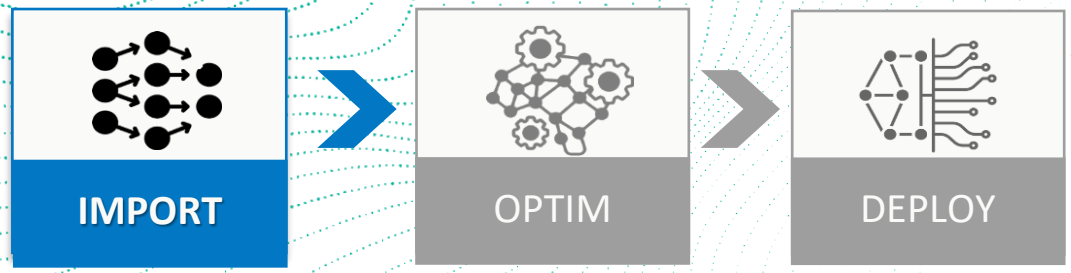
API

C++ + Python

Multi-Platform and Packaging

Windows Mac ubuntu PIP

Hosted by ECLIPSE FOUNDATION



aidge

- **High degree of interoperability** with ONNX standard

+60 operators and involved in the Safety ONNX standard

- Native support of the **main embedded architectures**

CNN, RNN, GAN, YOLO, Transformer and soon SNN

- Rich analysis tools to assess model complexity: parameters, operations, etc.

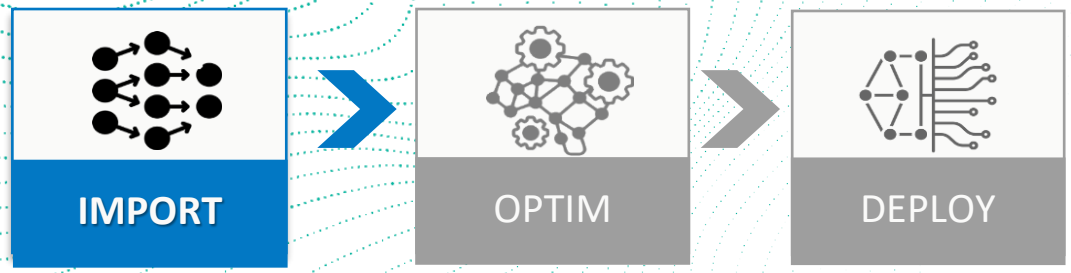
- Unique intermediate representation for easy model access and manipulation

ONNX coverage ratio
DINOv2 (Meta): 100%.

```
# Here show nb of operators!
aidge_onnx.native_coverage_report(dinov2_model)
```

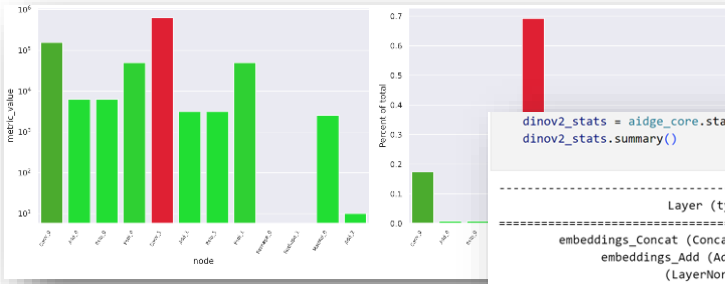
[4]

```
Native operators: 824 (17 types)
- Add: 159
- Concat: 1
- Conv2D: 1
- Div: 49
- Erf: 12
- Gather: 1
- MatMul: 72
- Mul: 73
- Pow: 25
- Producer: 209
- ReduceMean: 50
- Reshape: 49
- Softmax: 12
- Split: 12
- Sqrt: 25
- Sub: 25
- Transpose: 49
Generic operators: 0 (0 types)
Native types coverage: 100.0% (17/17)
Native operators coverage: 100.0% (824/824)
```



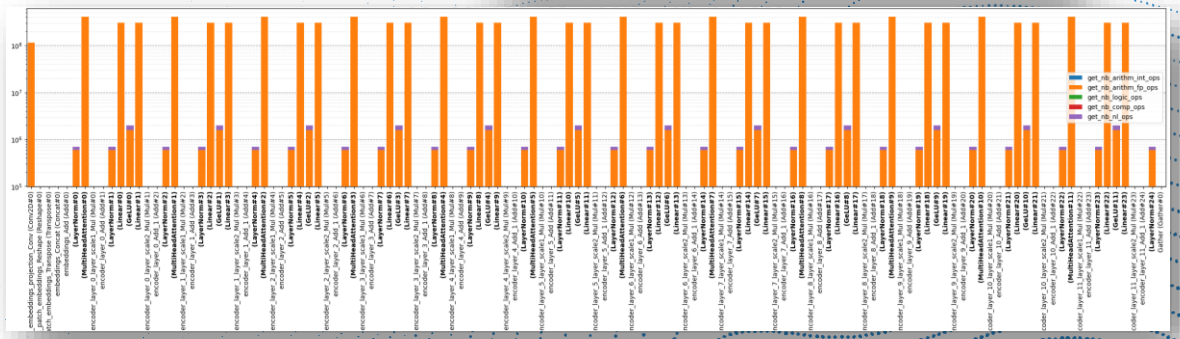
aidge

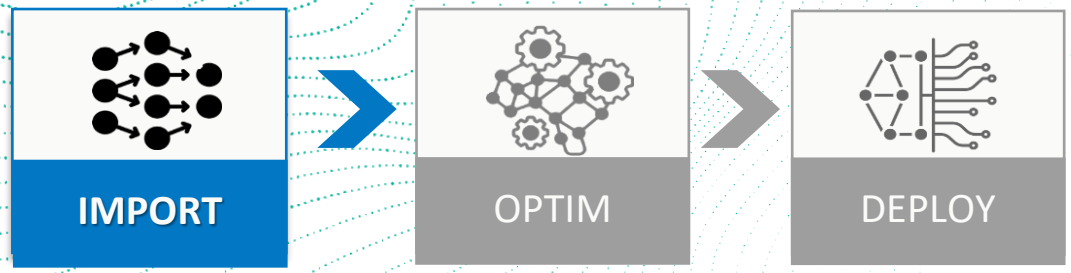
- High degree of interoperability with ONNX standard
- Native support of the main embedded architectures
- **Rich analysis tools to assess model complexity**: parameters, operations, etc.
- Unique intermediate representation for easy model access and manipulation




```
dinov2_stats = aidge_core.static_analysis.StaticAnalysisExt(dinov2_model)
dinov2_stats.summary()
```

```
----------------------------------------------------------------------------
                        Layer (type)               Output Shape         Param #
============================================================================
       embeddings_Concat (Concat#0)               [1, 257, 384]             384
            embeddings_Add (Add#0)                [1, 257, 384]           98688
                      (LayerNorm#0)               [1, 257, 384]             770
              (MultiHeadAttention#0)              [1, 257, 384]          591370
encoder_layer_0_layer_scale1_Mul (Mul#0)          [1, 257, 384]             384
         encoder_layer_0_Add (Add#1)              [1, 257, 384]               0
                      (LayerNorm#1)               [1, 257, 384]             770
                         (Linear#0)             [1, 257, 1536]          591360
                           (GeLU#0)             [1, 257, 1536]               3
                         (Linear#1)               [1, 257, 384]          590208
encoder_layer_0_layer_scale2_Mul (Mul#1)          [1, 257, 384]             384
       encoder_layer_0_Add_1 (Add#2)              [1, 257, 384]               0
                      (LayerNorm#2)               [1, 257, 384]             770
              (MultiHeadAttention#1)              [1, 257, 384]          591370
encoder_layer_1_layer_scale1_Mul (Mul#2)          [1, 257, 384]             384
         encoder_layer_1_Add (Add#3)              [1, 257, 384]               0
                      (LayerNorm#3)               [1, 257, 384]             770
                         (Linear#2)             [1, 257, 1536]          591360
                           (GeLU#1)             [1, 257, 1536]               3
...
Forward/backward pass size (MB): 74.537109375
Params size (MB): 82.50892639160156
Estimated Total Size (MB): 157.62025451660156
----------------------------------------------------------------------------
```

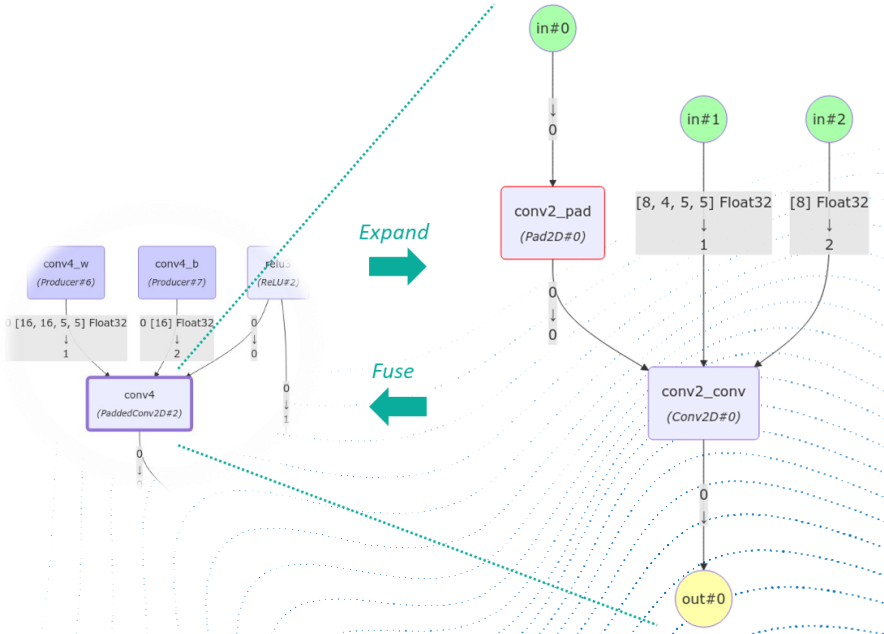




aidge

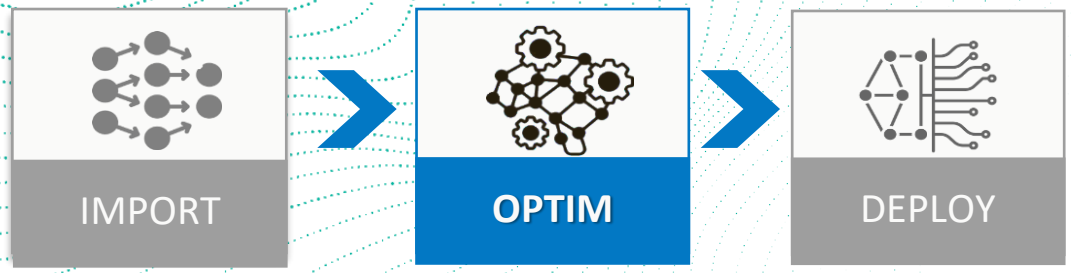
- High degree of interoperability with ONNX standard
- Native support of the main embedded architectures
- Rich analysis tools to assess model complexity: parameters, operations, etc.
- **Unique intermediate representation** for easy model access and manipulation

Match the granularity **required by the implementation**



A powerful **graph matching system**

```
aidge_core.fuse_to_metaops(dinov2_model,
   "ScaledDotProductAttention#1->Transpose->Reshape#1->Linear;"
   "Reshape#1<1~Producer;«
   "ScaledDotProductAttention#1<0-(Transpose<-Reshape#2<-Add#1);"
   "ScaledDotProductAttention#1<1-(Transpose<-Reshape#3<-Add#2);"
   "ScaledDotProductAttention#1<2-(Transpose<-Reshape#4<-Add#3);"
   "Reshape#2<1~Producer;, "MultiHeadAttention")
```

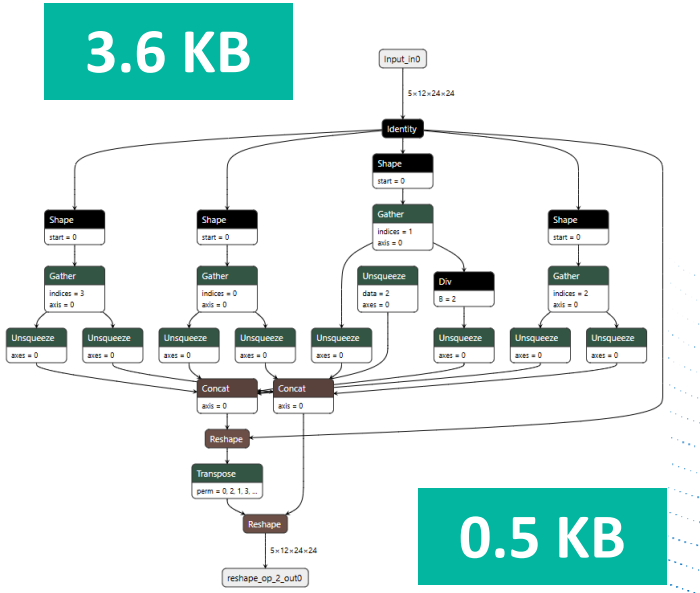


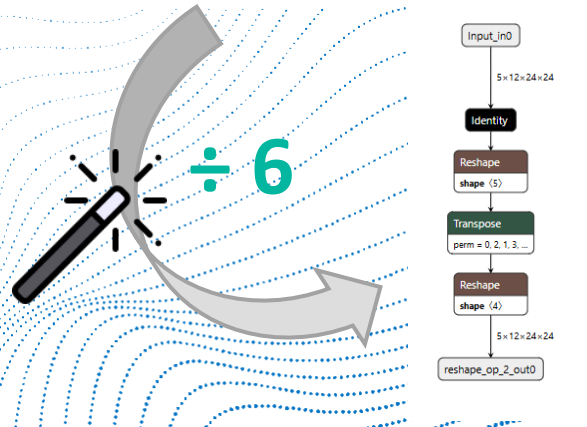
aidge

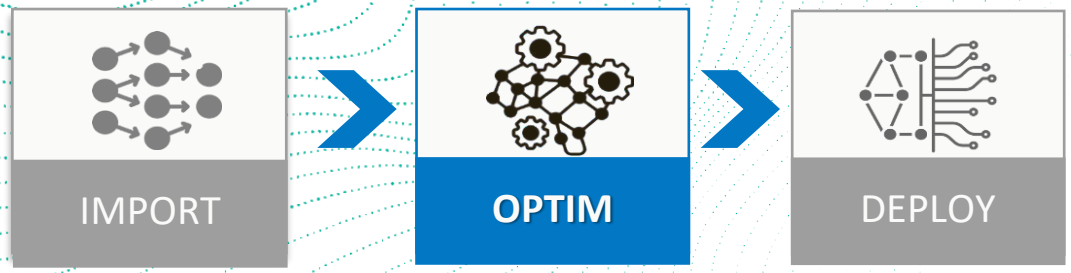
- **Automatic model reduction:** catalog of optimizations with deletion, reorganization and merging of operations

- **State-of-the-art quantification to desired accuracy (ResNet, VGG, etc.)**
  - After learning: without loss up to 8-bit integer
  - During learning: without loss up to 4-bit integer

- **Tensor decomposition compression method**
  - **ResNet-50 x ImageNet: 15% compression** without loss
  - **ResNet-18 x CIFAR 100: 45% compression** without loss

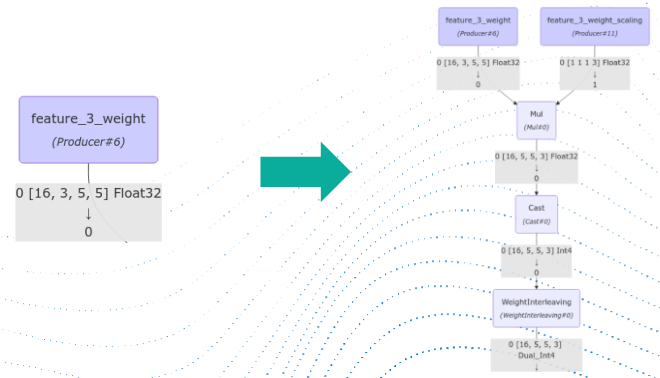




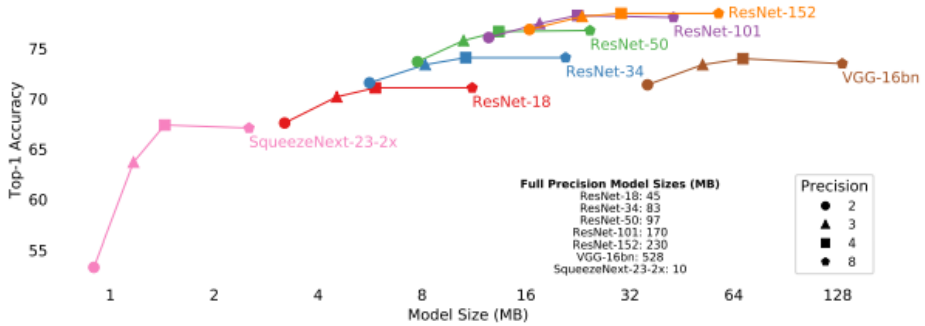


aidge

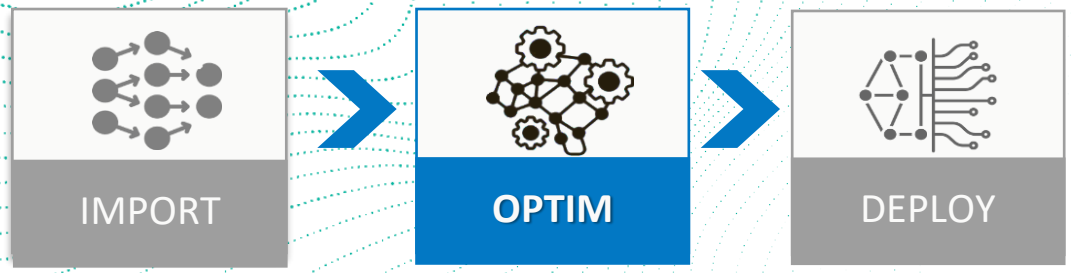
Traceability of optimization for certification purposes



- **Automatic model reduction:** catalog of optimizations with deletion, reorganization and merging of operations
- **State-of-the-art quantification to desired accuracy (ResNet, VGG, etc.)**
  - After learning: without loss up to 8-bit integer
  - During learning: without loss up to 4-bit integer
- **Tensor decomposition compression method**
  - **ResNet-50 x ImageNet: 15% compression** without loss
  - **ResNet-18 x CIFAR 100: 45% compression** without loss



LEARNED STEP SIZE QUANTIZATION, Esser et al.



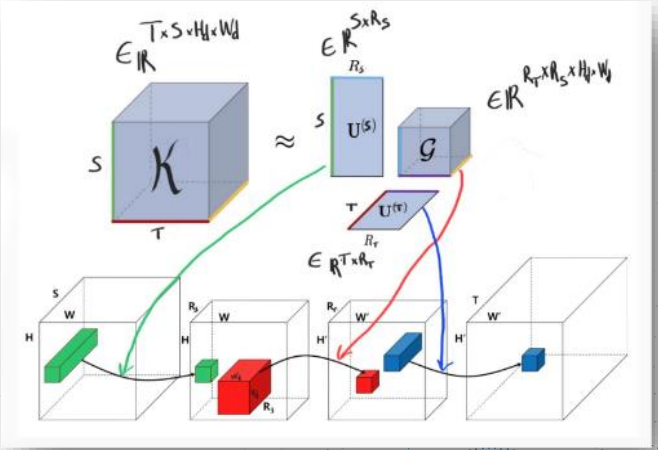
aidge

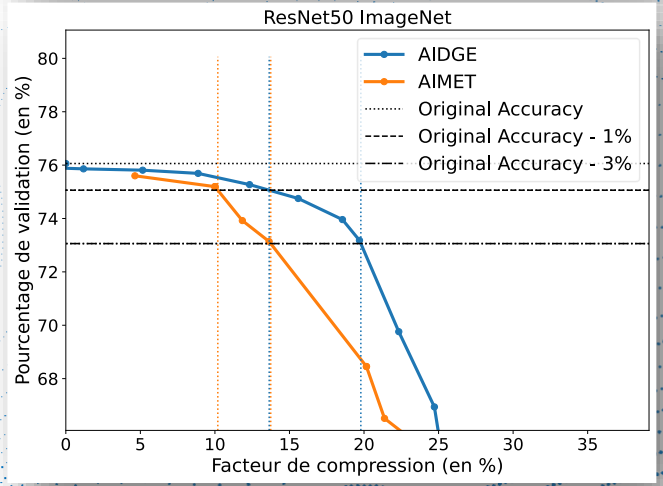
- **Automatic model reduction:** catalog of optimizations with deletion, reorganization and merging of operations

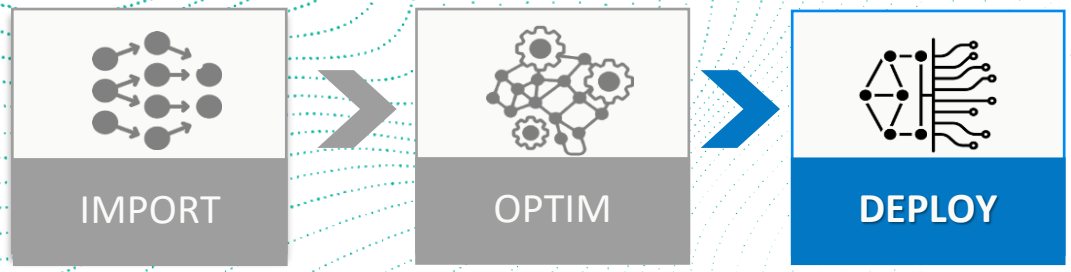
- **State-of-the-art quantification to desired accuracy (ResNet, VGG, etc.)**
  - After learning: without loss up to 8-bit integer
  - During learning: without loss up to 4-bit integer

- **Tensor decomposition compression method**
  - **ResNet-50 x ImageNet: 15% compression** without loss
  - **ResNet-18 x CIFAR 100: 45% compression** without loss

Higher performance than AIMET (Qualcomm)







aidge

- **ONNX export for interfacing with numerous SDKs**
- **Transparent, multi-paradigm code generation engine** (C/C++, HDL, etc.), enabling integration of compute kernels (native or third-party)
- **Multi-target reference export (C++) and specializations** (ARM, Texas Instrument SoC and ESP32 coming soon)
- **Orchestration control and memory optimization** through statistical allocation

STM32 Cube.AI

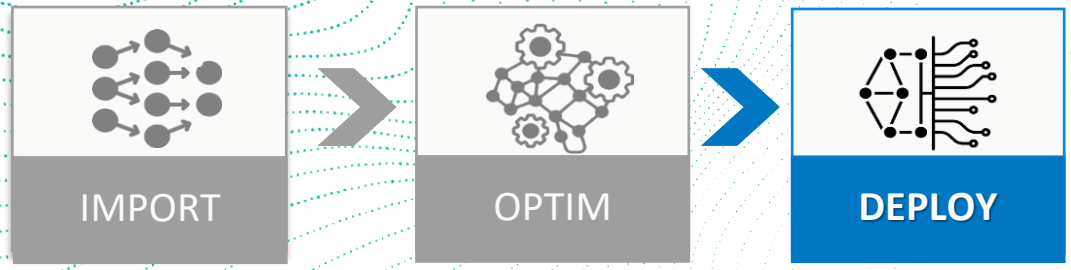
KaNN™ Kalray Neural Network

TensorRT

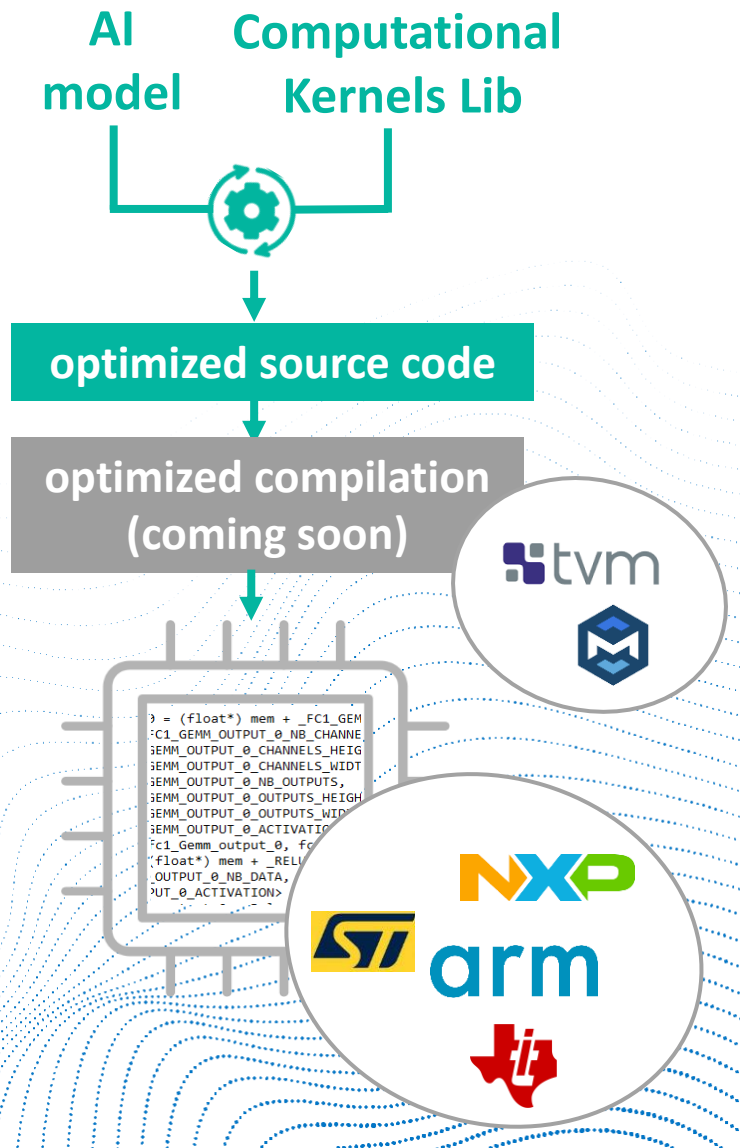
OpenVINO™

ONNX RUNTIME

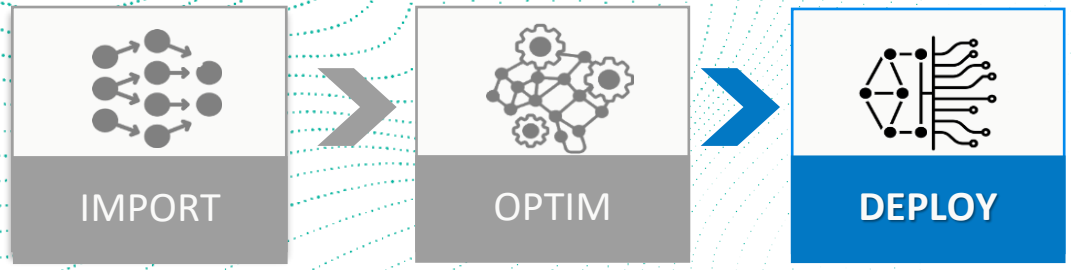
Qualcomm



aidge

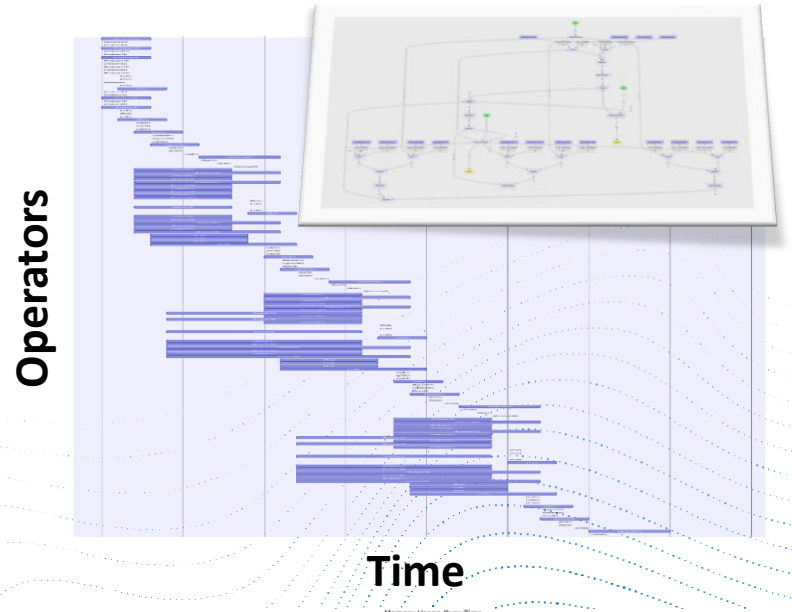


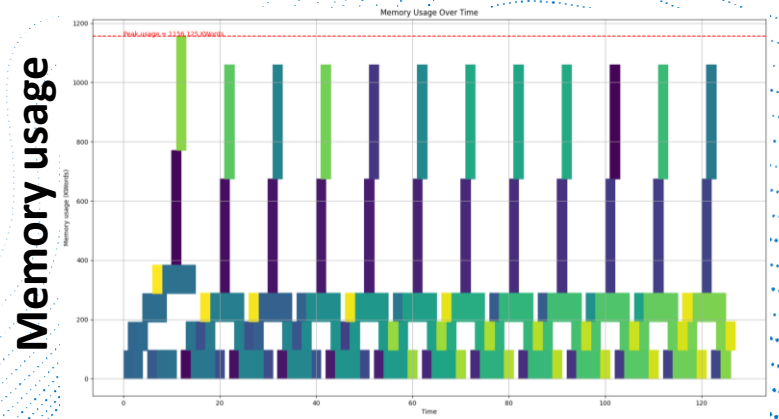
- **ONNX export for interfacing with numerous SDKs**
- **Transparent, multi-paradigm code generation engine** (C/C++, HDL, etc.), enabling integration of compute kernels (native or third-party)
- **Multi-target reference export (C++) and specializations** (ARM, Texas Instrument SoC and ESP32 coming soon)
  Coming soon :
  - Certification-aware export in C (ONERA)
  - Compilation workflow (INRIA)
- **Orchestration control and memory optimization** through statistical allocation



aidge

- **ONNX export for interfacing with numerous SDKs**
- **Transparent, multi-paradigm code generation engine** (C/C++, HDL, etc.), enabling integration of compute kernels (native or third-party)
- **Multi-target reference export (C++) and specializations** (ARM, Texas Instrument SoC and ESP32 coming soon)
- **Orchestration control and memory optimization** through statistical allocation





Active developments and collaborations

+40 submitters

5 organizations

+300 Commits /month

6th version

20cent fabricea wboussella diegob alalloyer oantoni jsimatic nvrlosemyself idealbuquerque clementgf julienl nthm
ikucher mnewson jeromeh marwaabd macario yberkat bobot alemesle thibaultallenet jgirardsatabin lucaslopez vlorrain
bhalimi gkubler raphaelmillet louislerbourg vbaudelet flebert axelfarr cguillon silvanosky hleborgne operrin mick94
obichler pineapple lrakotoarivony sylvainbataille mszczep farnez na25 lsoulier cmoineau noamzerah alicebatte hrouis

Norms

EUROCAE afnor ONNX

+30 industrials partners

THALES (Building a future we can all trust) infineon AIRBUS KALRAY ALSTOM MBDA (Missile Systems) Valeo ArcelorMittal NX NanoXplore ST EDF SAFRAN KNDS ...

+10 academic partners

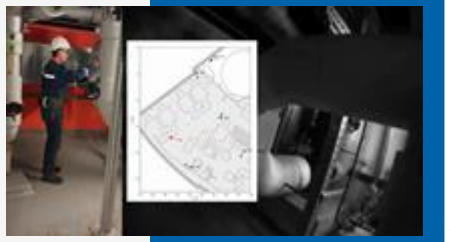
SAL Silicon Austria Labs Fraunhofer cea Inria ONERA (The French Aerospace Lab) cnrs université Paris-Saclay ...

Some use cases

Defect detection and classification

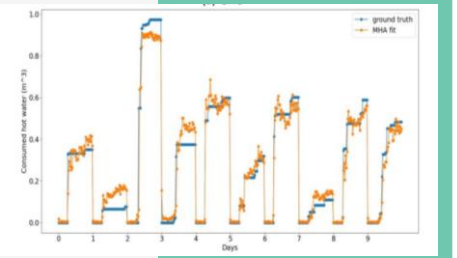
- **Low latency (20m/s) and high performance algorithm to detect small defect (~mm) with low contrast**
- Deployment on Nvidia GPU
- In collaboration with ArcelorMittal

Indoor localisation using multi-sensors



- **Lightweight prediction algorithm based on IMU sensors combined with fast AI Visual tracking (x15 faster)**
- Deployment on STM32
- In collaboration with EDF sysnav

Heat Pump Monitoring



- **Lightweight prediction algorithm based on incremental learning** for adaptive heat pump control and monitoring
- Deployment on STM32
- Up to 40% energy saving
- In collaboration with EDF UGA Université Grenoble Alpes

Hardware design : NeuroCorgi AI-ASIC accelerator

- **RTL generation of quantized model**
- HD images processing in real time : latency is less than 10ms
- Uses 1,000 times less power than commercial circuits cea

aidge

Follow us

Join us and chat !