DOLPHIN DESIGN

TECHNOLOGY PATENTED

TinyRAPTOR : AI NPU* ENGINE

WHAT IS TinyRAPTOR ?

TinyRAPTOR is a programmable hardware accelerator that specializes in neural network (NN) inference and vision processing. It includes 128 processing elements and a buit-in DMA. Designed for a broad spectrum of low-power AI applications, it efficiently executes deep neural networks (DNNs). Raptor's fully programmable nature allows it to be tailored to meet the specific requirements of any application.
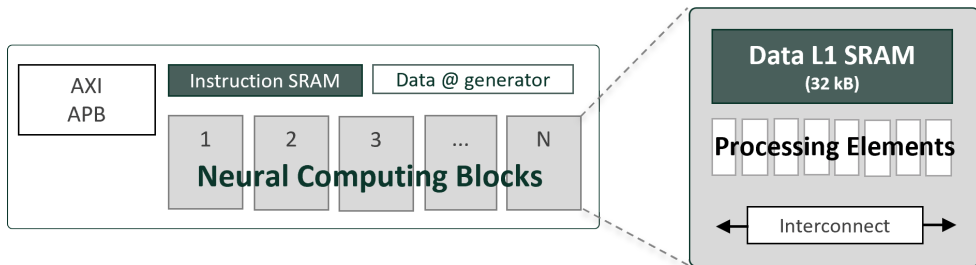
Additionally, it is fully interoperable with other SPEED platforms, including SPIDER, CHAMELEON, PANTHER, and BAT, creating a comprehensive AI development platform. This integration ensures accelerated time-to-market and distinct market differentiation for users.
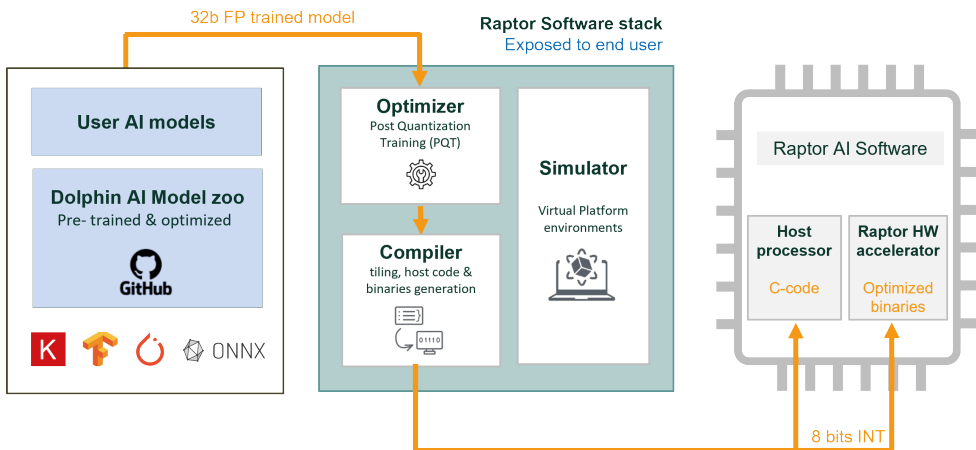
APPLICATIONS

- Face recognition
- Fire alarm detection
- Asset counting
- AR/VR
- Battery powered AI devices
- Robotics and Drones

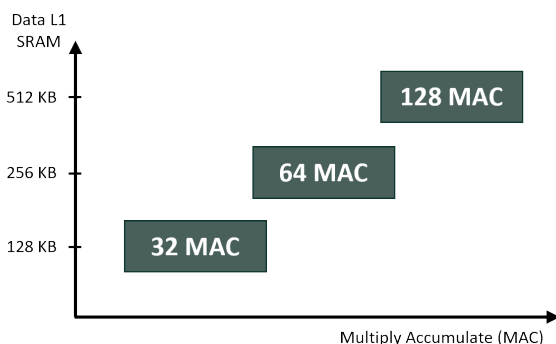*Note: NPU stands for Neural Processing Unit

TinyRAPTOR NPU HW block

AXI APB

Instruction SRAM | Data @ generator

1 | 2 | 3 | ... | N

Neural Computing Blocks

Data L1 SRAM (32 kB)

Processing Elements

Interconnect

TinyRAPTOR SDK

32b FP trained model

Raptor Software stack
Exposed to end user

User AI models

Dolphin AI Model zoo
Pre- trained & optimized

GitHub

K | ONNX

Optimizer
Post Quantization Training (PQT)

Simulator

Virtual Platform environments

Compiler
tiling, host code & binaries generation

Raptor AI Software

Host processor
C-code

Raptor HW accelerator
Optimized binaries

8 bits INT

Key Features

- Near Memory Computing architecture

- Up to 128 MAC/Cycle

- 20 inferences/second (<1mW)

- 92% Top-1 accuracy

- 73% $mAP_{0.5}$ accuracy

- Small memory footprint

TinyRAPTOR Portfolio

Data L1 SRAM

512 KB

256 KB

128 KB

128 MAC

64 MAC

32 MAC

Multiply Accumulate (MAC)

DOLPHIN
DESIGN

www.dolphin-design.fr/tinyraptor