
brainchip™

brainchip™

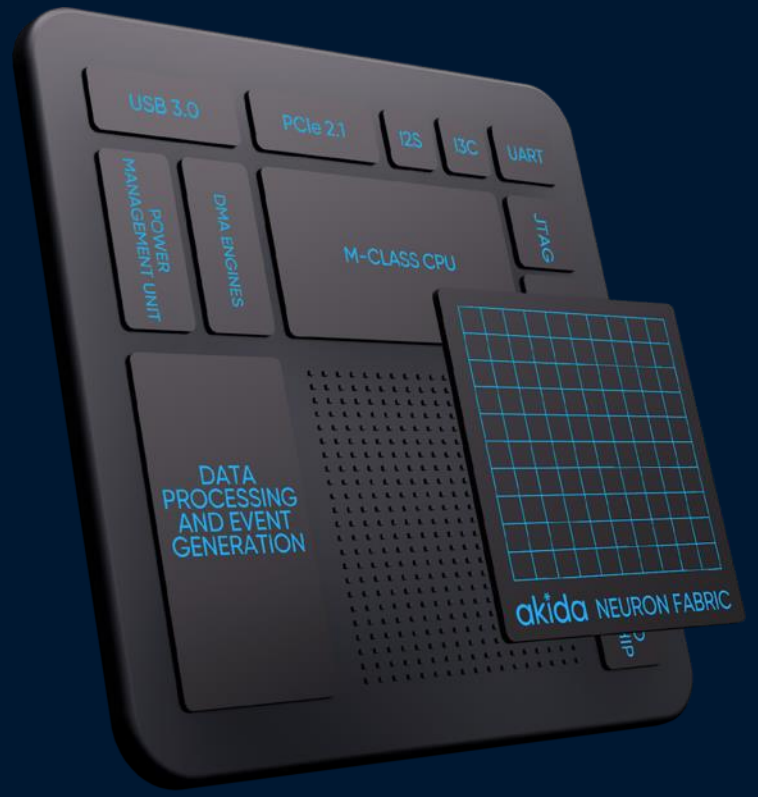
Akida™

A Hardware Accelerator for AI tasks, for integration with RISC-V Systems

Douglas McLelland[1] and Gregor Lenz[2]

Contact: dmclelland@brainchip.com

1. BrainChip, Toulouse, France
2. Neurobus, Toulouse, France



NEUROBUS

brainchip™

Overview

- Akida: AI accelerator to complement RISC-V
- A simple use case
- Hardware Performance

Akida

A Hardware Accelerator for AI Workloads

- Akida IP is integrated into chip designs
- Configure size to target workload
- Event-based Processing
- At-Memory Compute
- Physical chips available for evaluation
- GRAIN: RISC-V + Akida is coming soon!

Akida Evaluation Hardware Devices

Akida 1.0

AKD1000 Chip

- Embedded M.4 Core
- 300 MHz
- 20 Nodes = 80 NPUs

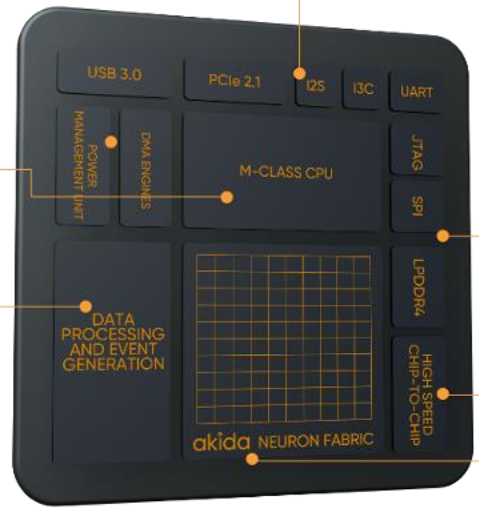
AKD1500 Chip

- PCIe Co-processor
- 400 MHz
- 8 Nodes = 32 NPUs

Data Input Interfaces
- PCI Express 2.1 x2 Lane Endpoint
- USB 3.0 Endpoint
- I3S, I2C, UART, JTAG

On-Chip Processor
- M-Class CPU with FPU & DSP
- System Management
- Akida Configuration

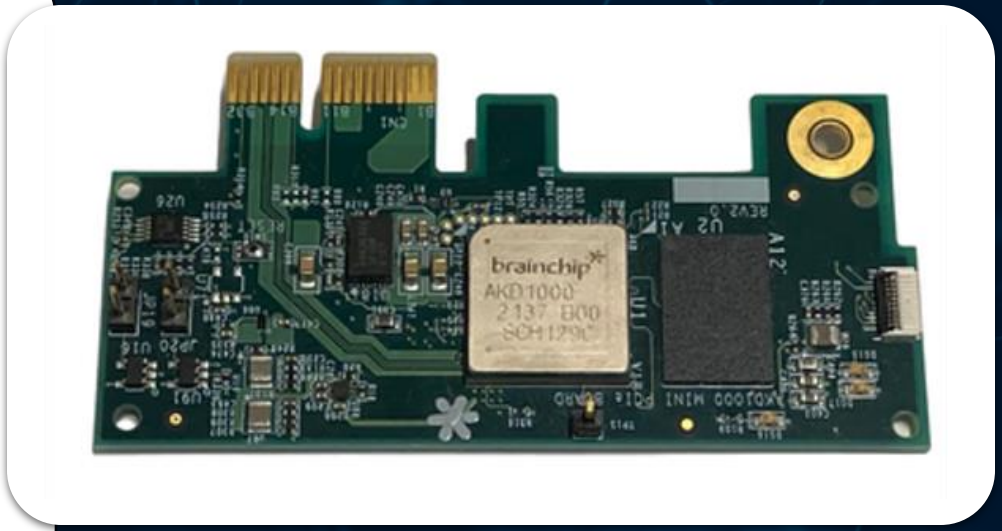
Data Processing
- Pixel-Event Converter
- SW Data-Event Encoder
- Any multivariable digital data
- Sound, pressure, temp., others



External Memory Interfaces
- SPI FLASH for boot/storage
- LPDDR4 Program/Weights

Multi-Chip Expansion
- PCIe 2.1 2 lane root complex
- Connects up to 64 devices

Flexible Akida Neuron Fabric
- Implements 80 NPUs
- All Digital logic with SRAM (8MB)
- Also Available as Licensed IP Core
- First Implementation: TSMC 28nm

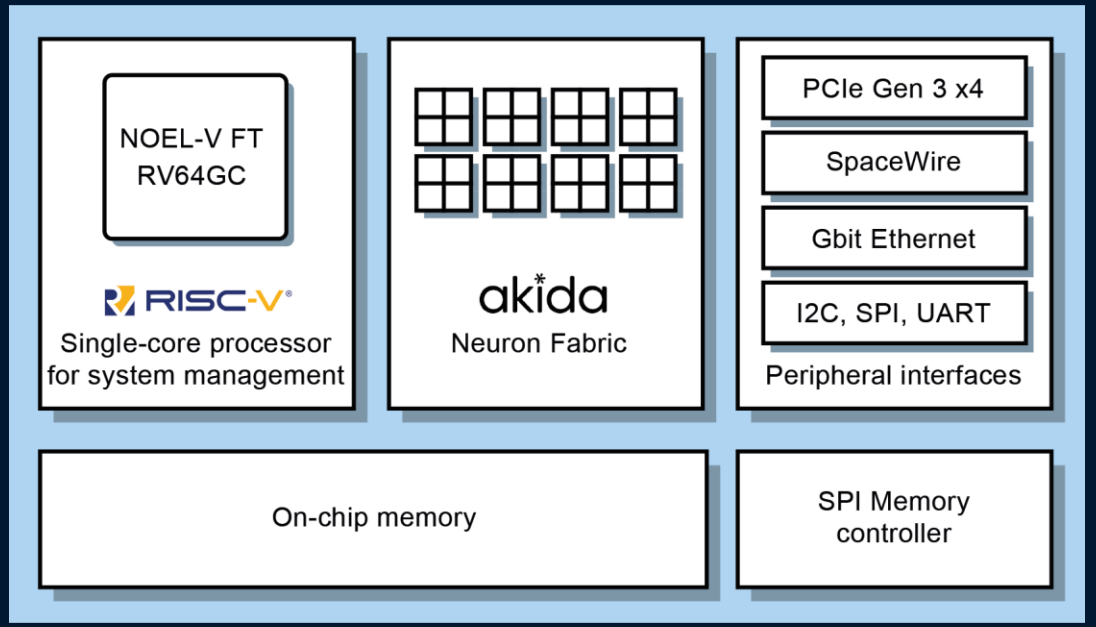


brainchip

FRONTGRADE Gaisler

GRAIN: a RISC-V + Akida Device is On Its Way!

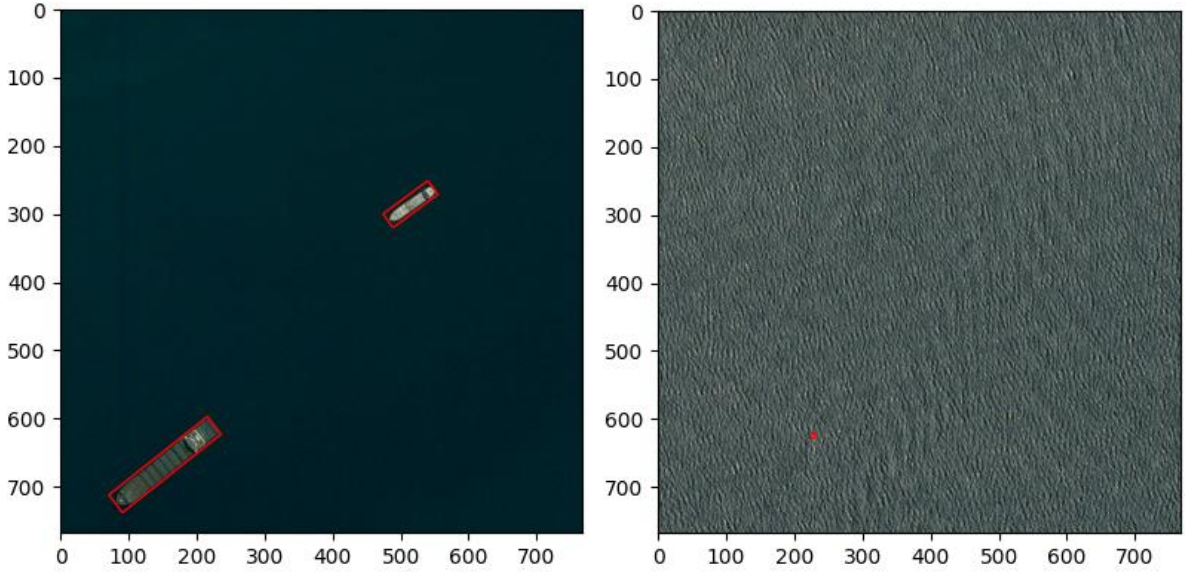
- RISC-V CPU (NOEL-V)
- Akida engine
- Fault-tolerant
- Rich set of interfaces allowing for versatile applications
- On-Chip RAM
- Small footprint

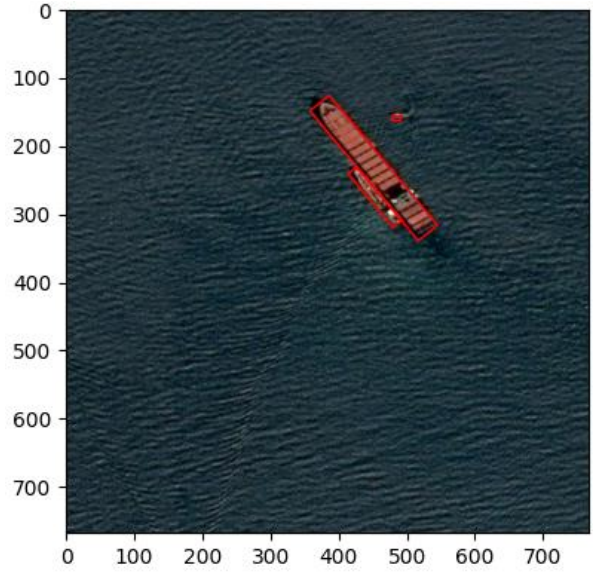


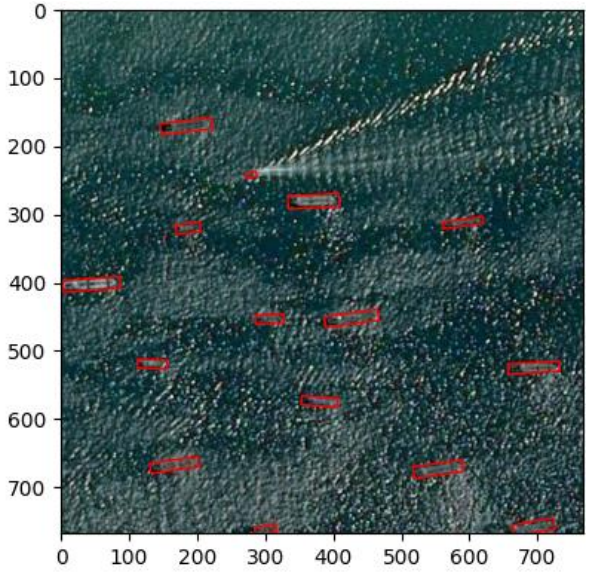
Ship Detection

A Simple Illustrative Use-Case

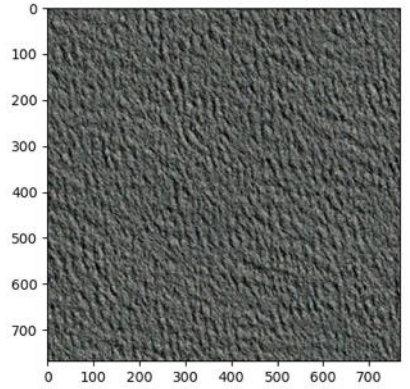
brainchip Airbus Ship Detection Dataset

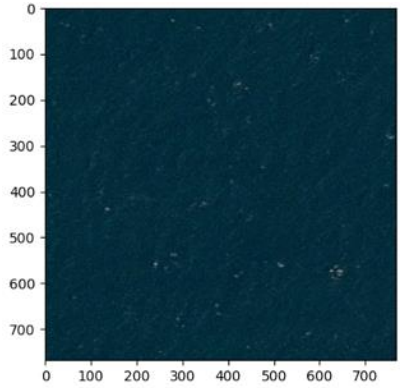


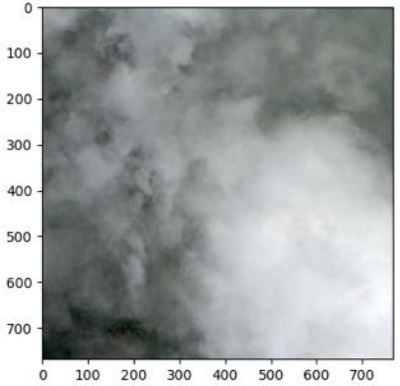




RGB image size	768 × 768
Total number of images	192,555
Number of training images	154,044
Percentage of images that contain ships	22.1%
Total number of bounding boxes	81,723
Median diagonal of all bounding boxes	43.19px
Ratio of bounding box to image area	0.3%







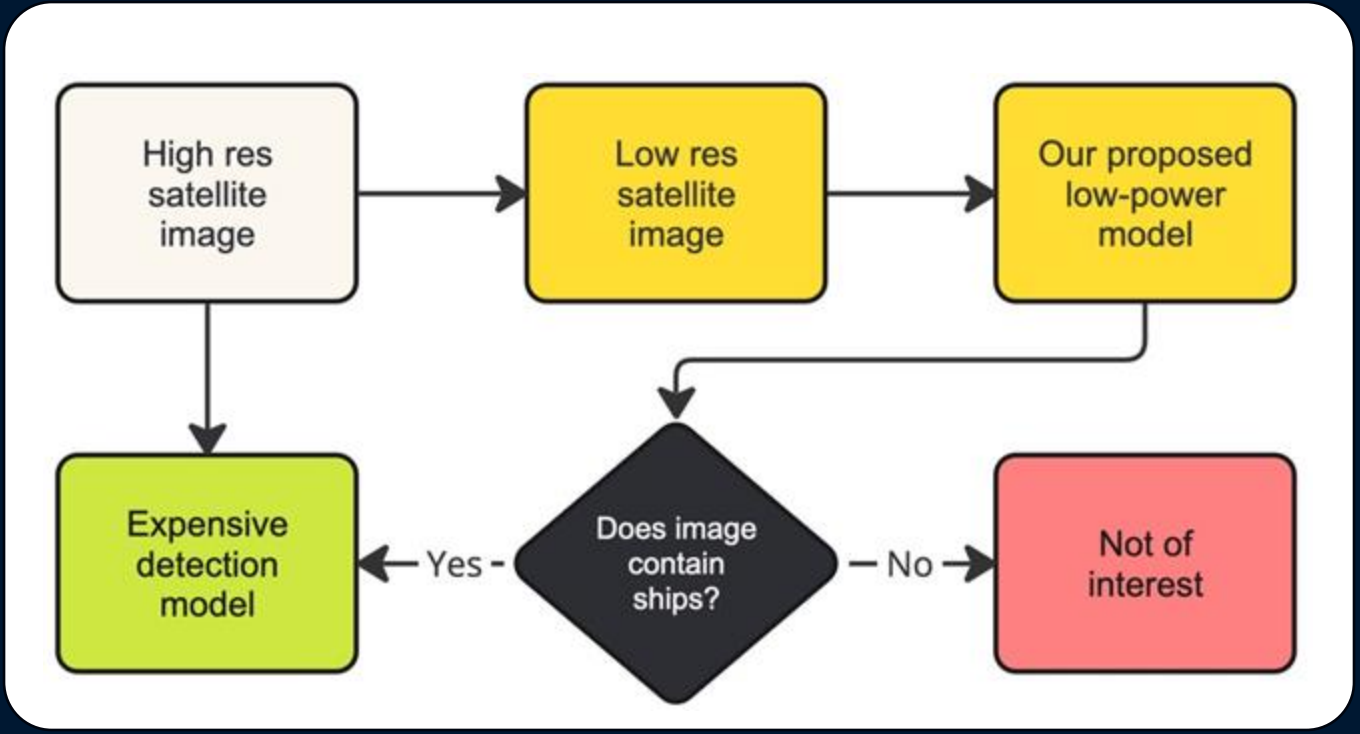
Detection Pipeline

Detection (spatial localization of objects) requires high resolution and larger models, high energy

- e.g. *YOLOv5m running on NVIDIA Jetson Nano, **2.9J per image** (7.81 W at 2.7 fps) (Machado et al, 2022)

Classification (*e.g.* ship/no-ship) can be done with high performance at lower resolution, with smaller models

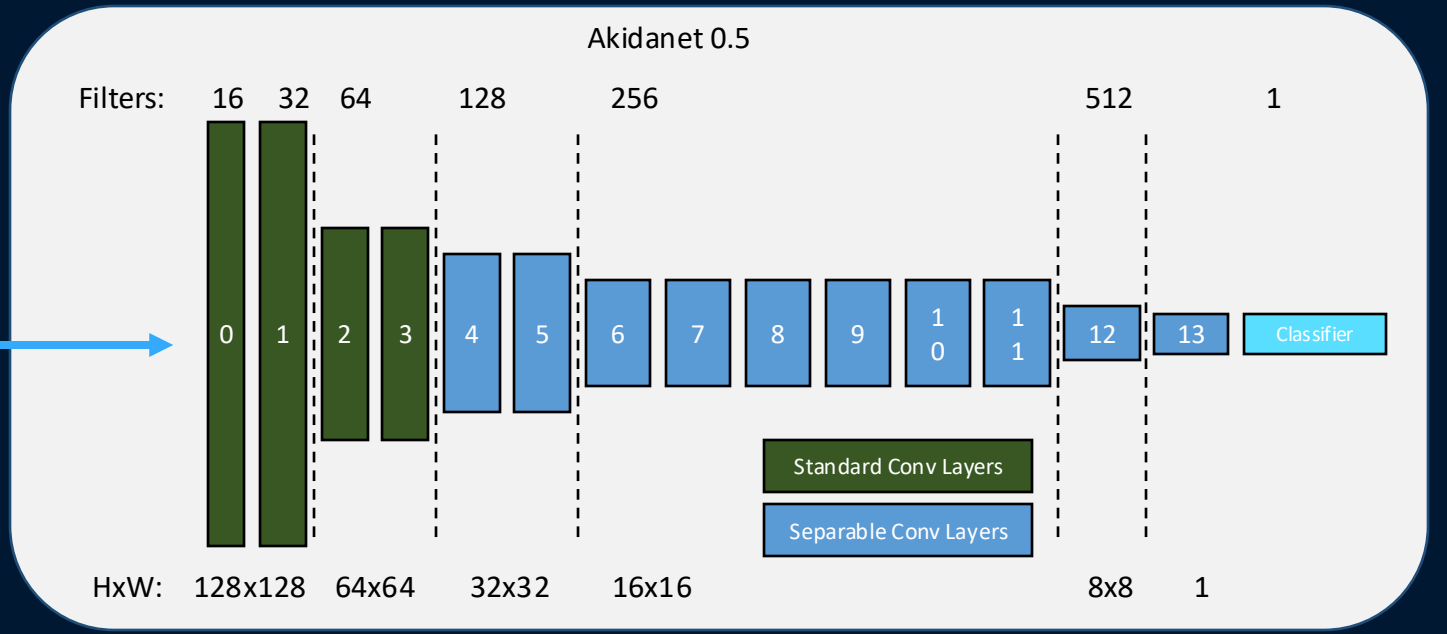
Efficient **system**:



Ship/No-Ship Classification Model

- Based on MobileNet v1 α=0.5 (Howard et al, 2017)
- 866k parameters
- Convolutional architecture
- 256x256 pixel input resolution
- Model trained with standard keras pipeline
- Quantized to 4-bit (weights and activations)

	Full Precision	4-bit	QAT	Recall "Bias"
Accuracy	97.91	95.75	97.67	
Recall	95.23	85.12	94.40	97.64
Precision	95.38	95.32	95.07	89.73



Hardware Performance

- Simple Inference Acceleration
- Increase acceleration using more NPs
- Process batch data using serial pipeline
- Handle larger models with multiple passes

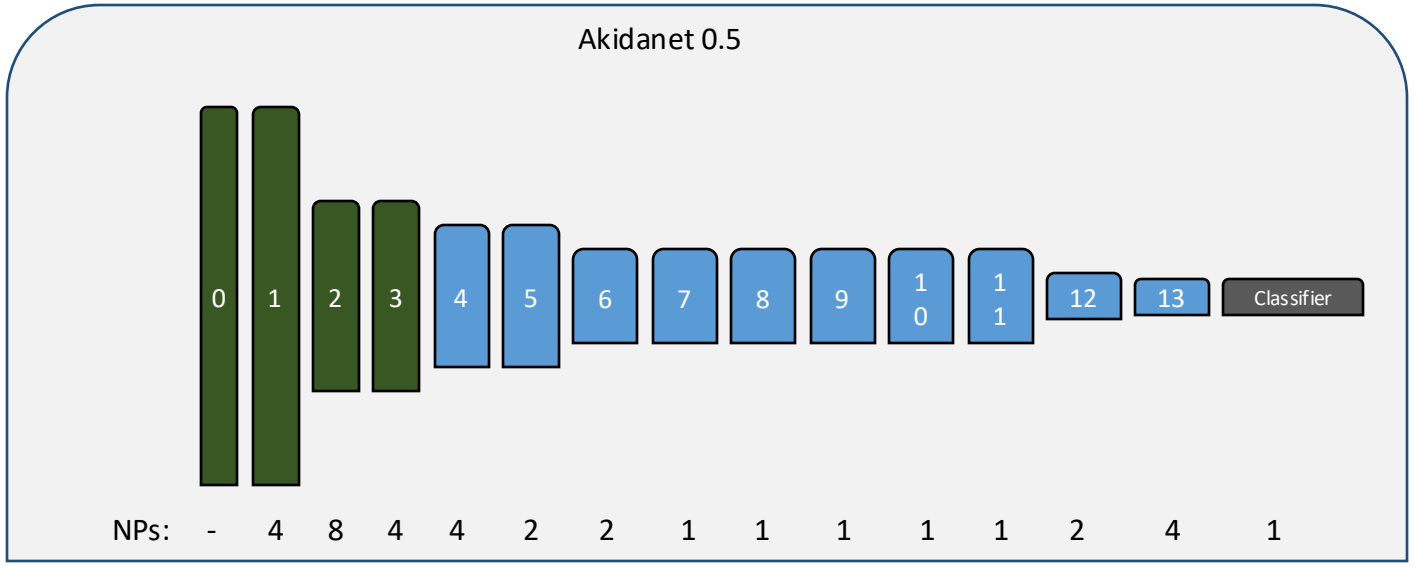
brainchip

Simple Inference:
Processing a single image

Hardware Device: AKD1000 (80 NPs available)

Model Summary

Input shape	Output shape	Sequences	Layers	NPs
[256, 256, 3]	[1, 1, 1]	1	15	36

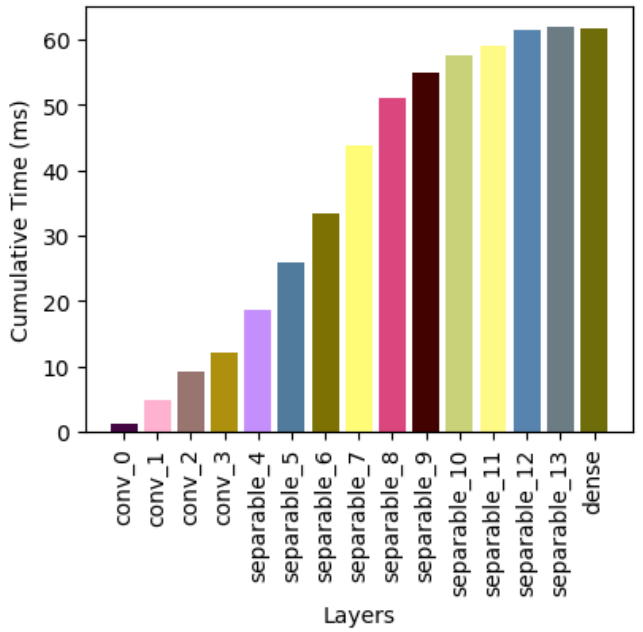


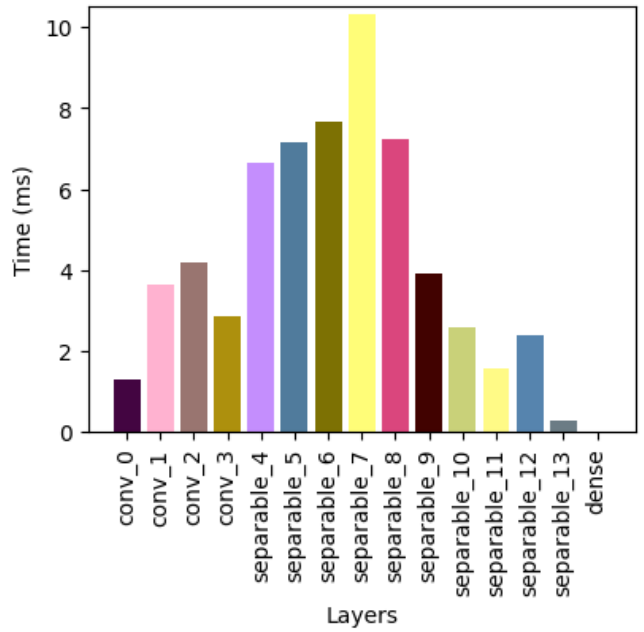
Layer (type)	Output shape	Kernel shape	NPs
======= HW/conv_0-dense (Hardware) - size: 881760 bytes ======			
conv_0 (InputConv.)	[128, 128, 16]	(3, 3, 3, 16)	N/A
conv_1 (Conv.)	[128, 128, 32]	(3, 3, 16, 32)	4
conv_2 (Conv.)	[64, 64, 64]	(3, 3, 32, 64)	8
conv_3 (Conv.)	[64, 64, 64]	(3, 3, 64, 64)	4
...			
separable_12 (Sep.Conv.)	[8, 8, 512]	(3, 3, 256, 1)	2
		(1, 1, 256, 512)	
separable_13 (Sep.Conv.)	[1, 1, 512]	(3, 3, 512, 1)	4
		(1, 1, 512, 512)	
dense (Fully.)	[1, 1, 1]	(1, 1, 512, 1)	1

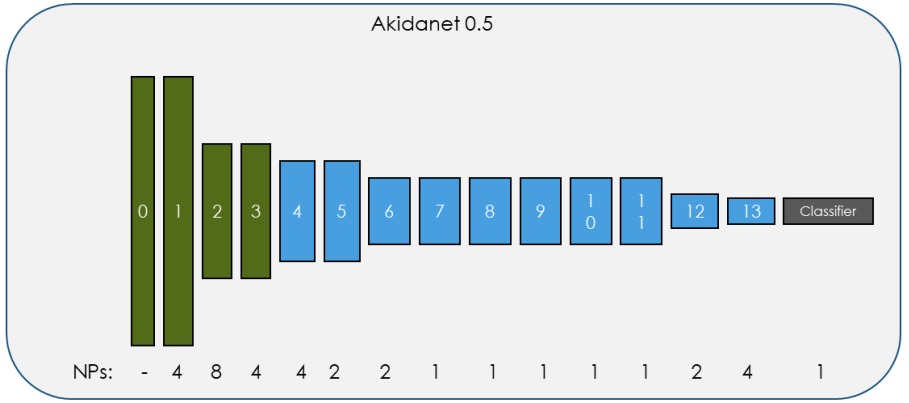
Simple Inference: Processing a single image

Hardware Device: AKD1000 (80 NPs available)

- Measured Processing Clock ticks
- Results show time for a 400 MHz device
- Per layer results measured by adding layers one by one



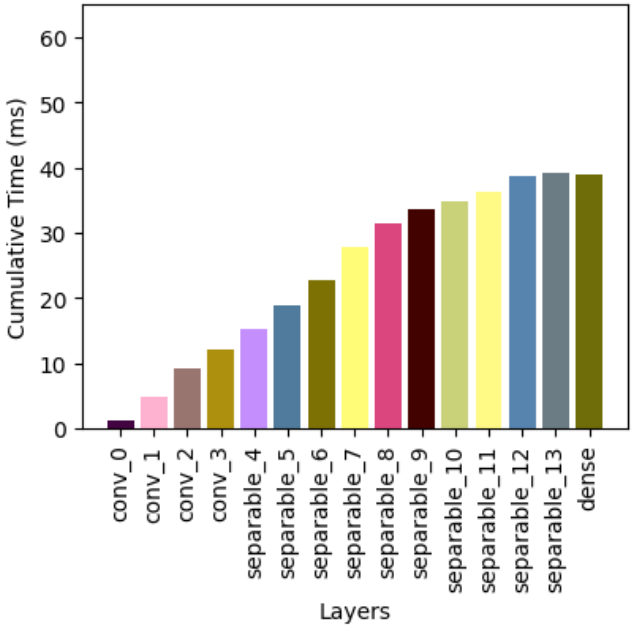


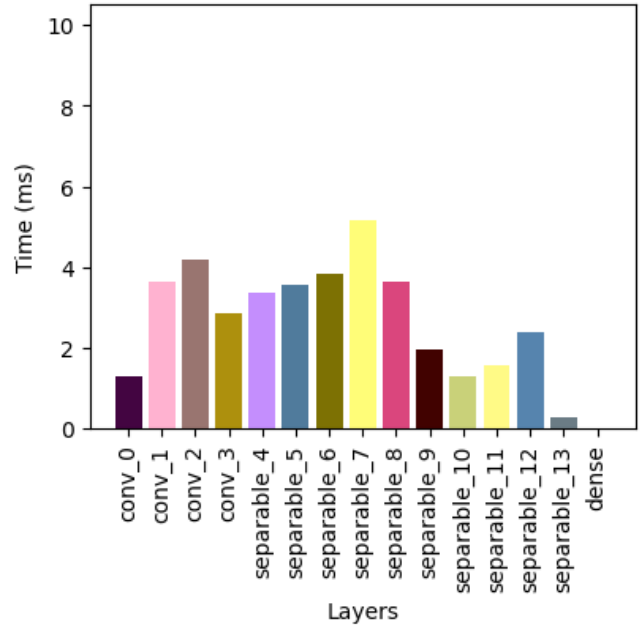


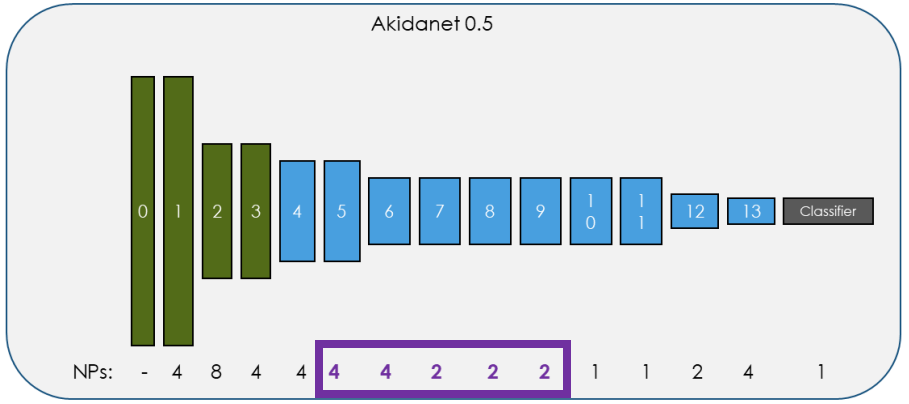
Simple Inference: Processing a single image

Hardware Device: AKD1000 (80 NPs available)

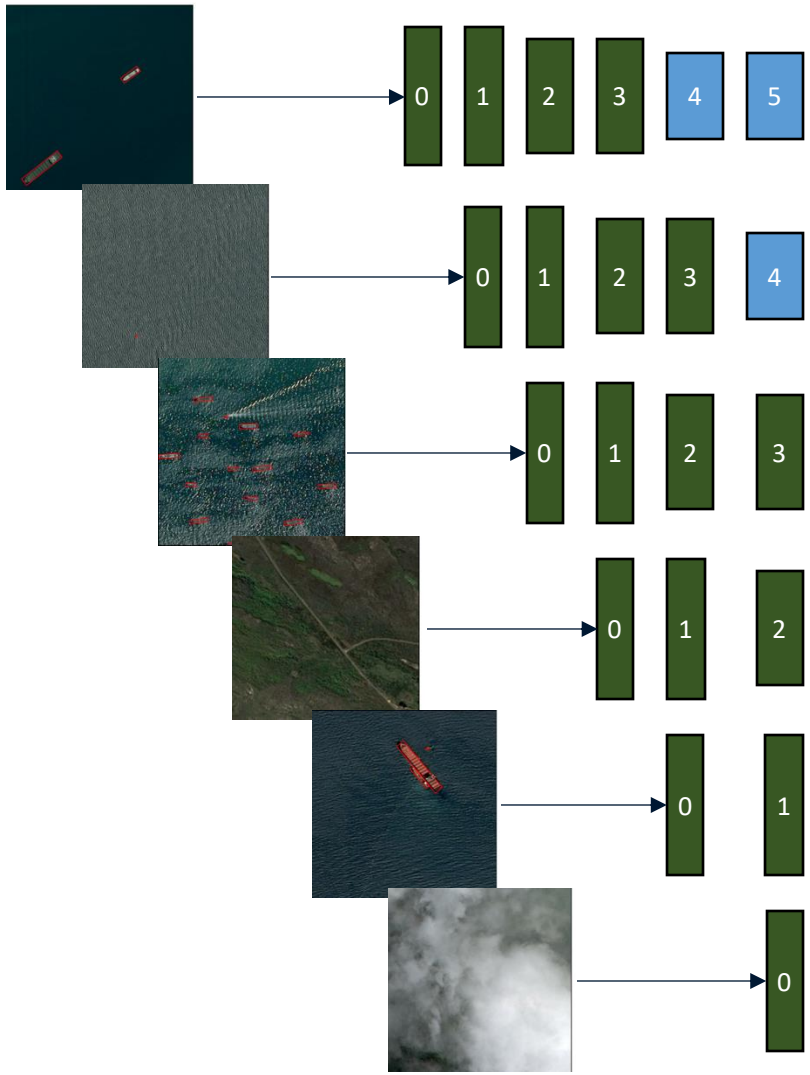
- Allocate more NPs to the middle layers

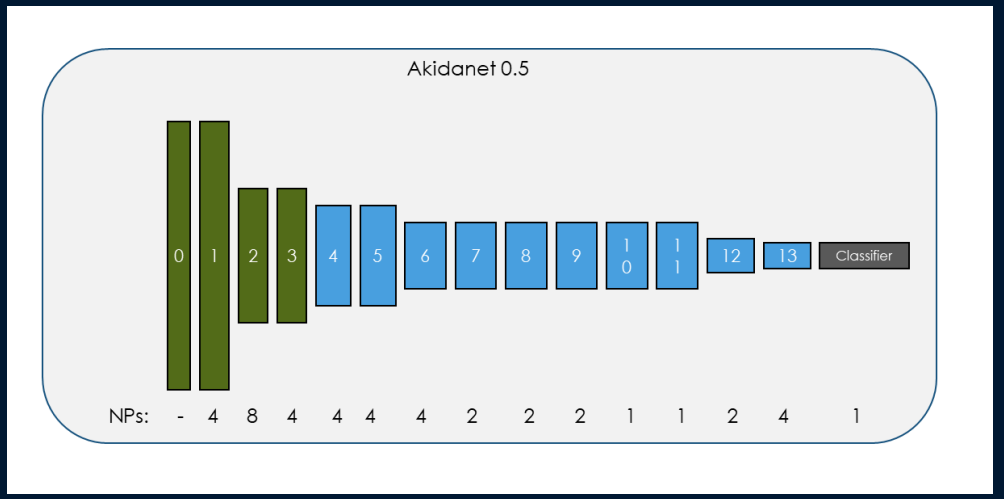






Pipeline Mode: Processing a batch of images





Pipeline Mode: Processing a batch of images

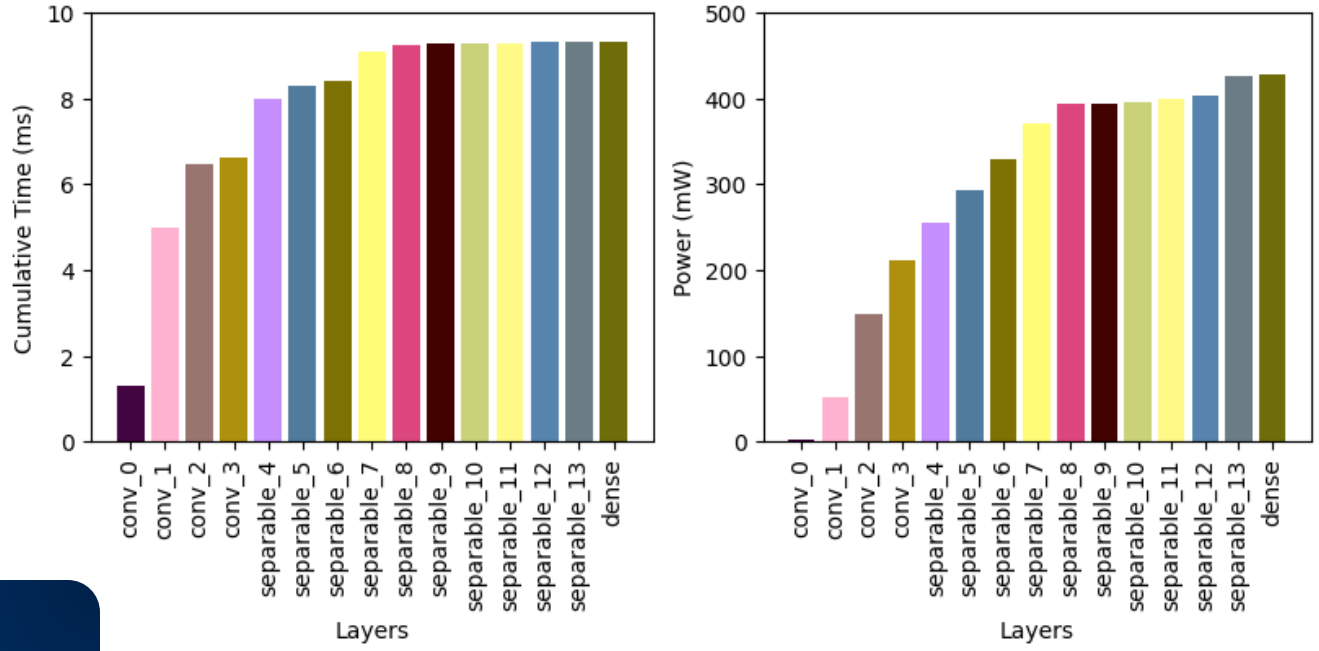
Hardware Device: AKD1000
(80 NPs available)

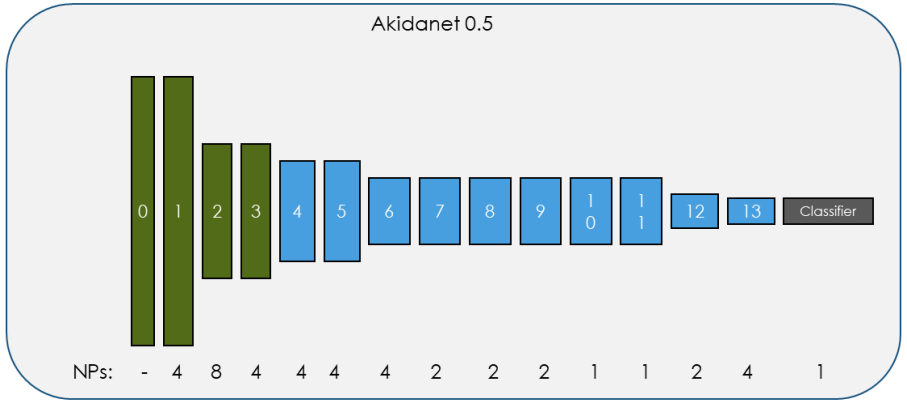
Akidanet 0.5 Model:

Time: 9.3 ms

Power: 428 mW

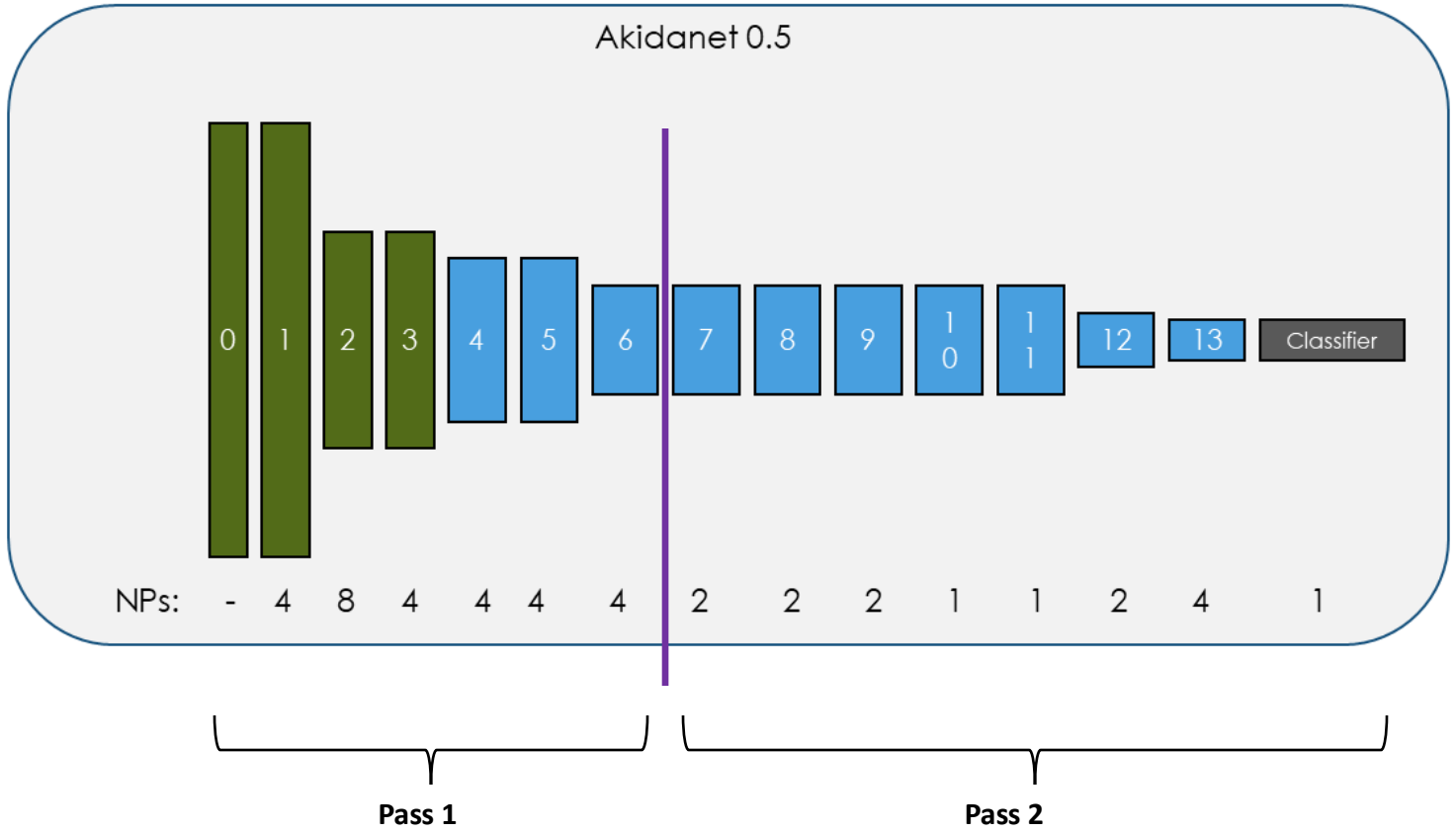
Energy: 4.0 mJ/frame





Multi-pass Inference: When Model is too Big to fit in memory

Hardware Device: AKD1500 (32 NPs available)

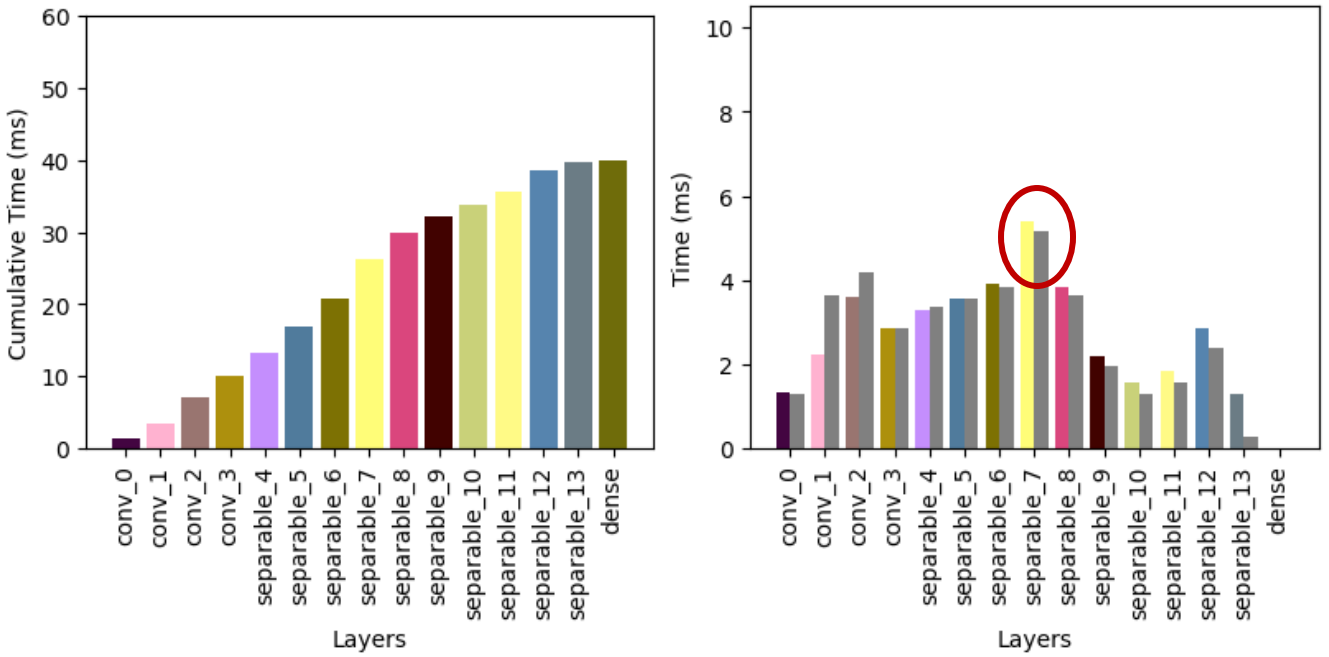


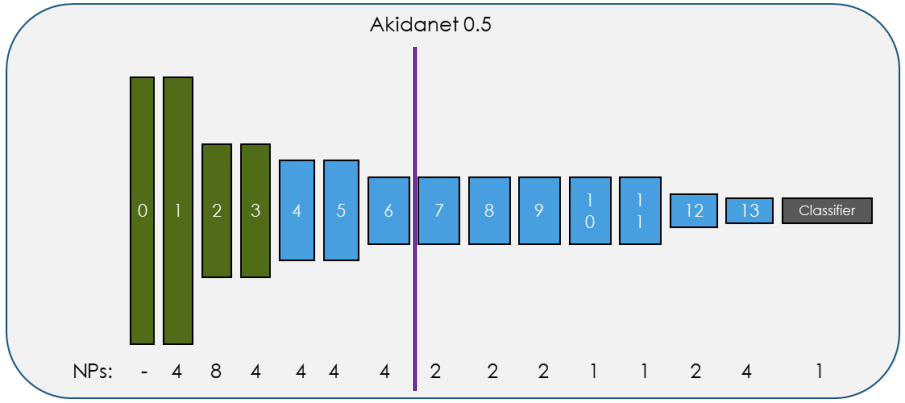

```
...
_________________________________________________________________
separable_5 (Sep.Conv.)   [32, 32, 128]   (3, 3, 128, 1)    4
_________________________________________________________________
                                (1, 1, 128, 128)
_________________________________________________________________
separable_6 (Sep.Conv.)   [16, 16, 256]   (3, 3, 128, 1)    4
_________________________________________________________________
                                (1, 1, 128, 256)
============ pass 2 ====================
separable_7 (Sep.Conv.)   [16, 16, 256]   (3, 3, 256, 1)    2
_________________________________________________________________
                                (1, 1, 256, 256)
_________________________________________________________________
separable_8 (Sep.Conv.)  [16, 16, 256]  (3, 3, 256, 1)  2
_________________________________________________________________
...
```

Multi-pass Inference: When My Model is too Big

Hardware Device: AKD1500 (32 NPs available)

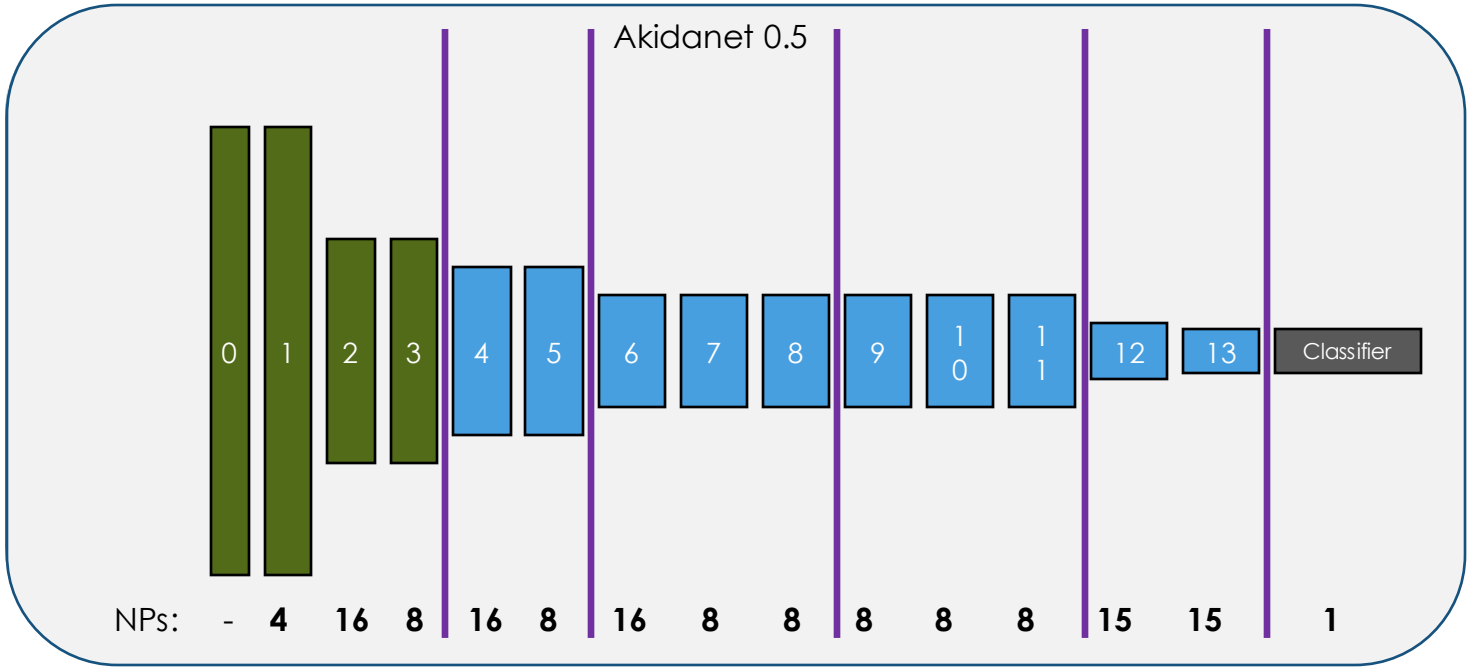
- Greatly increased flexibility to handle larger models
- At a cost of increased bandwidth (weights loaded every frame)
- ... but intermediate values still stay entirely on-chip





Multi-pass Inference: Maximise Throughput

Hardware Device: AKD1500 (32 NPs available)

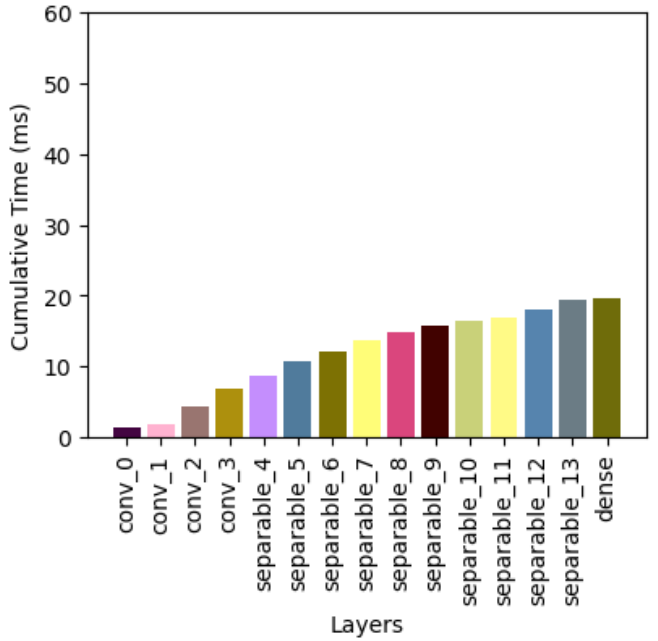


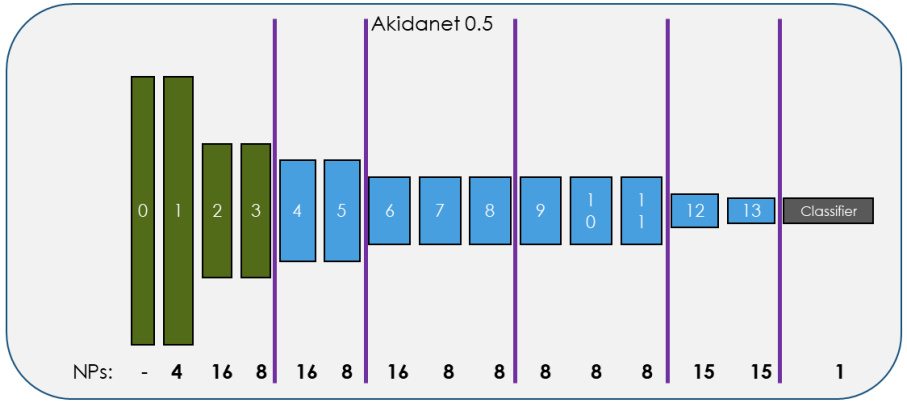
Model Summary

Input shape	Output shape	Sequences	Layers	NPs
[256, 256, 3]	[1, 1, 1]	1	15	**139**

Multi-pass Inference: Maximise Throughput

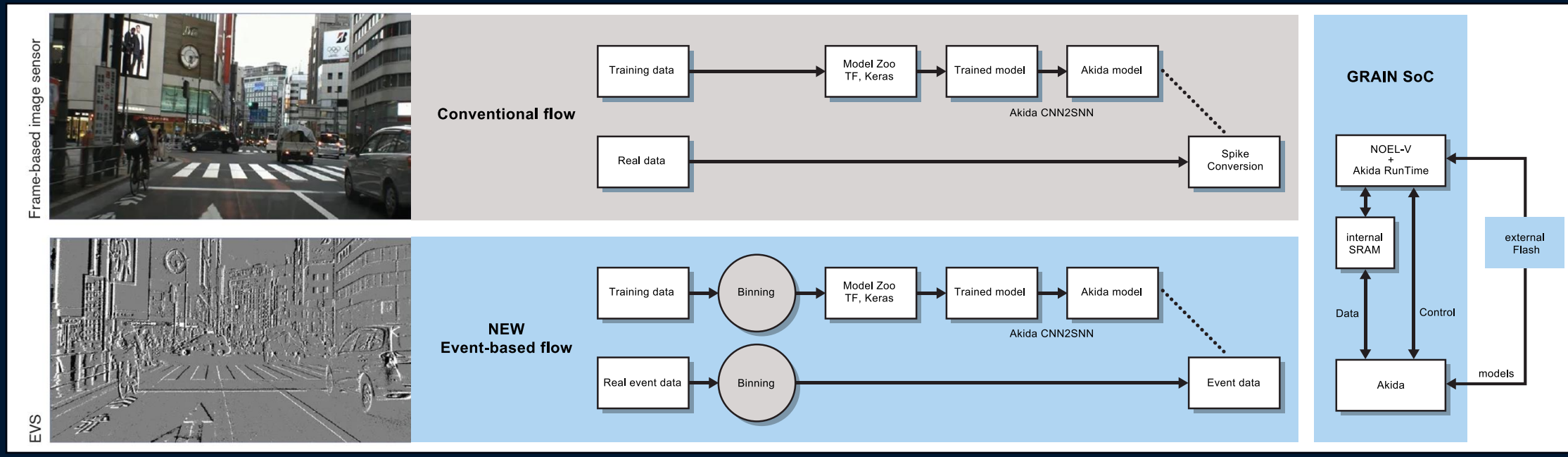
Hardware Device: AKD1500
(32 NPs available)





Event-based Processing

- Poster:



Conclusion

The Power of Processing

- Akida – fast, low-power AI acceleration
- Integration with RISC-V: GRAIN chip from Frontgrade Gaisler
- Highly configurable for a range of constraints

Acknowledgements

Gregor Lenz @ Neurobus

Thank You

Get It Touch with

brainchip™

Website

Https://www.brainchip.com

Email

dmclelland@brainchip.com

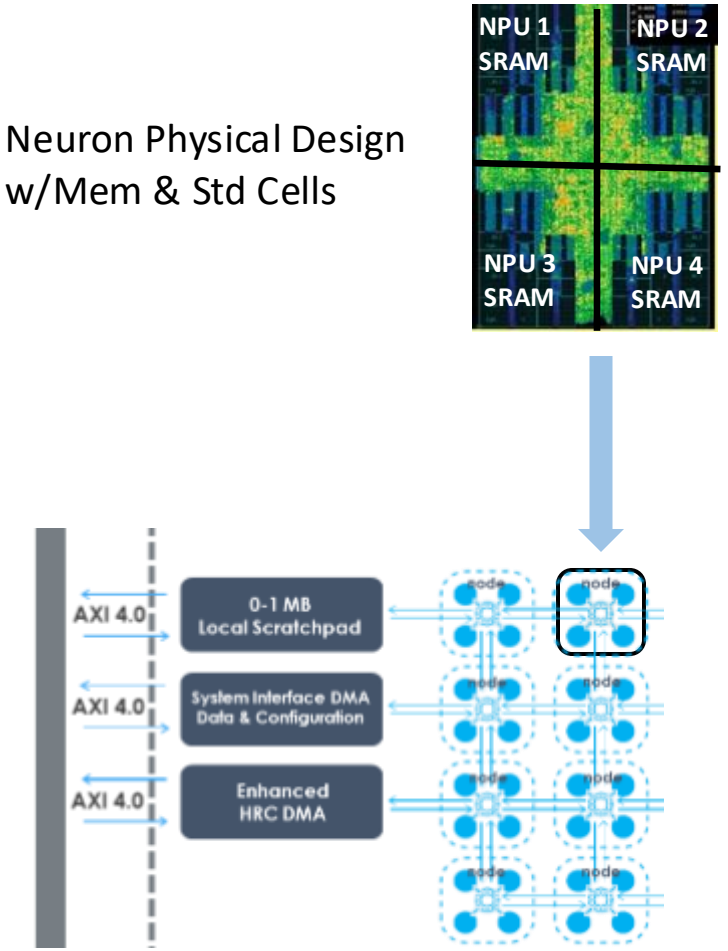
brainchip

Akida IP

What's in a Neuron?

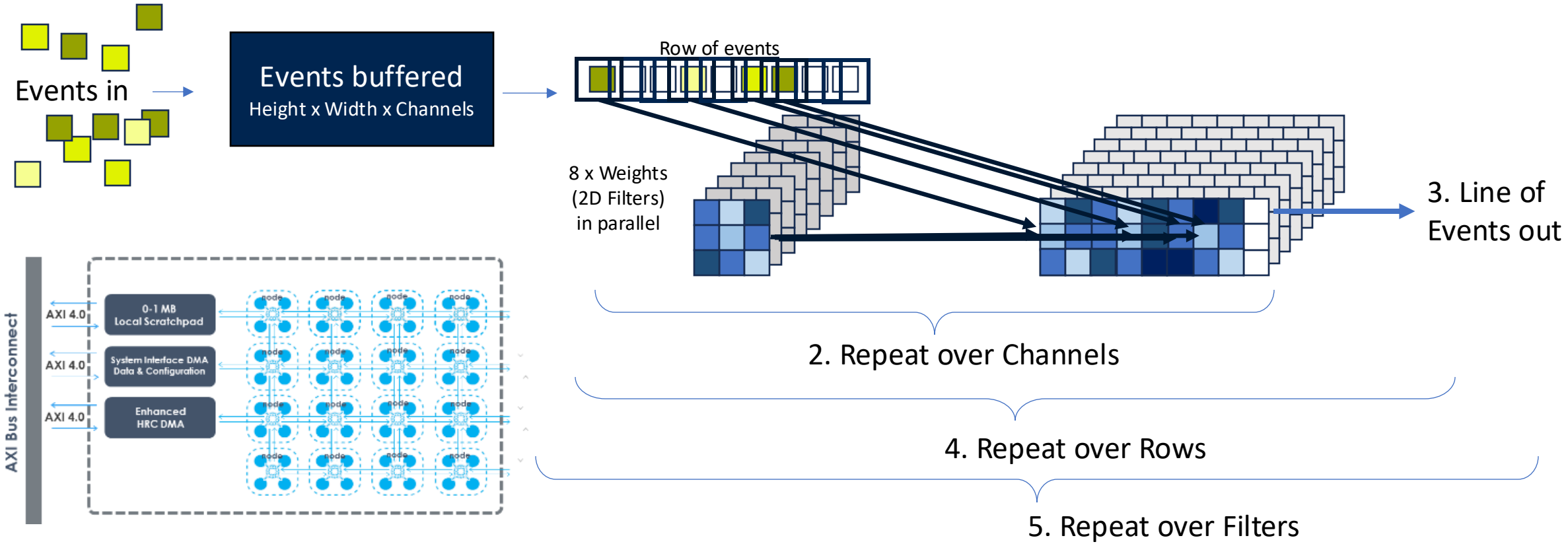
- Single Node has 4 NPU engines
- One NPU configurable up to 100KB at compute memory
- Standard Configuration is 100KB/NPU (400K per neural node)
- All NPU engines run on a single clock implementation
- On-Chip communication via mesh network
- Runtime SW manages configurations
- Physical Implementation enables tiled expansion of network
- Scalability and consistency between Process Technologies

Neuron Physical Design
w/Mem & Std Cells



Event-based Processing in Akida

What's in an Akida Neural Processor?

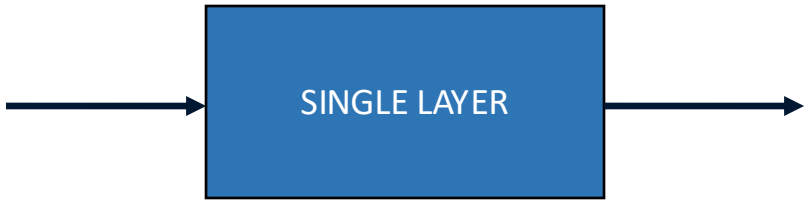


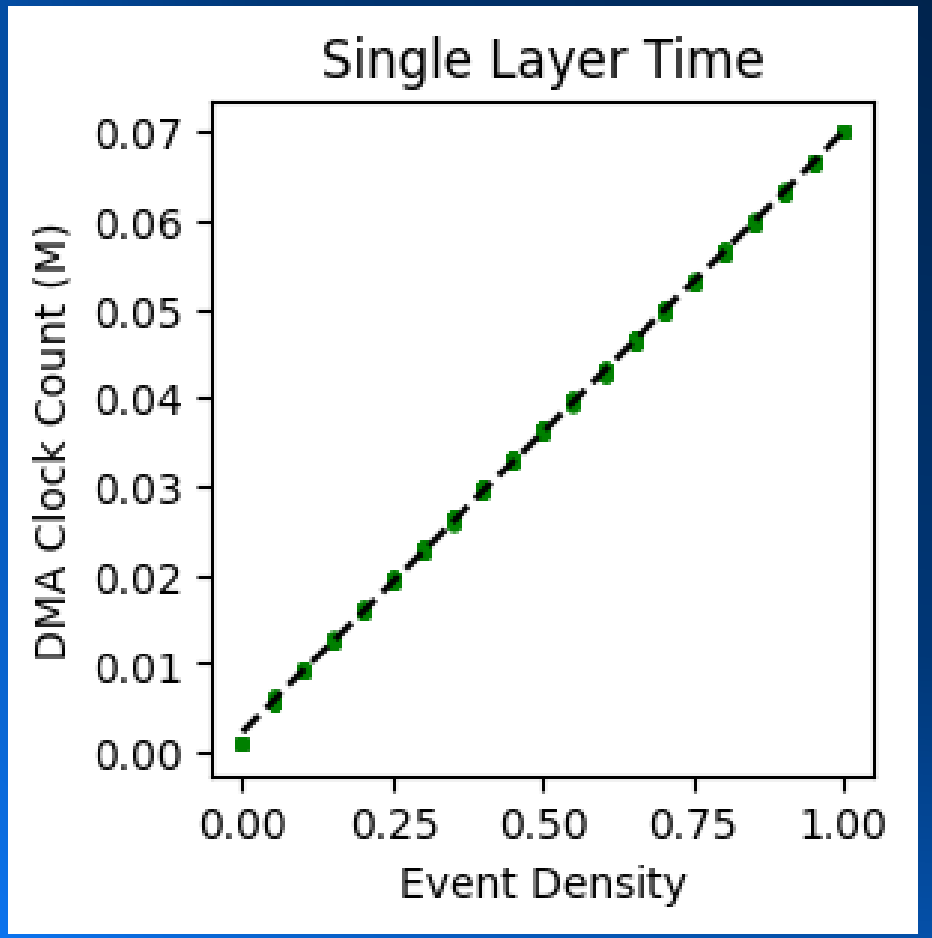
Event-based Processing in Akida

Processing Time vs “Sparsity”

EVENT DENSITY

10% 50% 90%





Sparsity is not just for Event-Based Inputs!

Sparsity for the Akidanet 0.5 / Ship-classification Model