HERONIC

Automating the Design of Bespoke AI Hardware

Imperial College London

Si SILICON CATALYST

UKRI Innovate UK

AI SuperConnector

altera solution acceleration partner

the problem

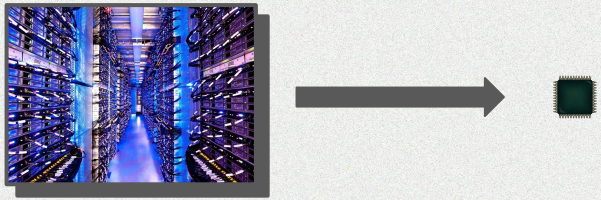
AI is costing the world

Datacenters are predicted to reach **1000 TWh** (equivalent to Japan) by 2030.

Is there a better way to run AI?

Edge AI

Running AI models on smaller, energy-efficient devices.



email: info@heronic.ai

the problem

Edge AI hardware has a utilisation problem

Potential

2000
TOPS

Reality

35 TOPS
(2% Utilisation)

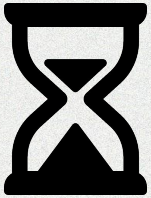
(Leading MLPerf Edge benchmark submission)

The **potential** for AI hardware has been rising YoY.

But the **actual** utilisation is **diminishing**, leading to:

Low Energy Efficiency

Poor Performance



Expensive AI

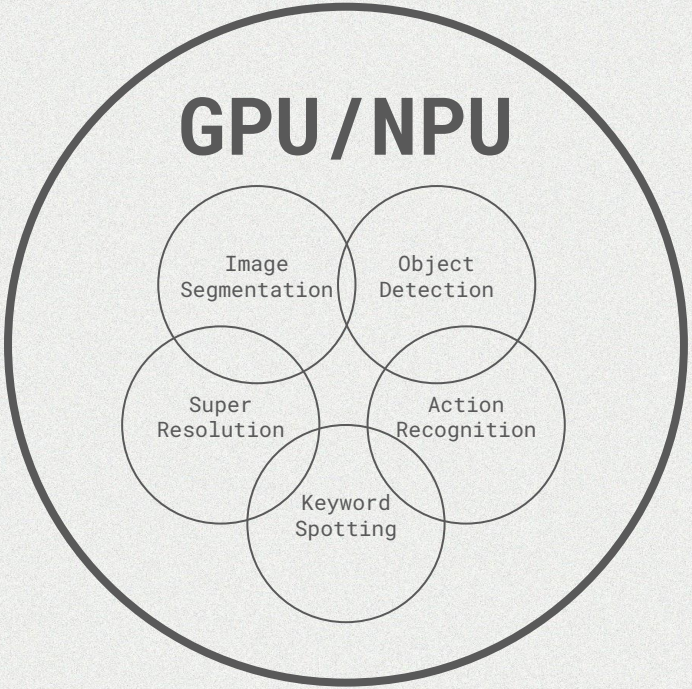
email: info@heronic.ai

the solution

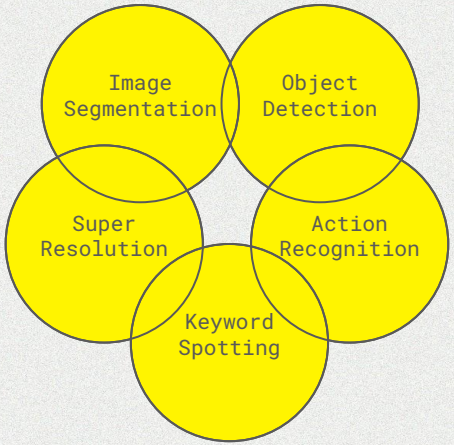
To address poor utilisation, Heronic is leveraging **AI** and **automation** is the key to finding the best **architecture** for AI acceleration

email: info@heronic.ai

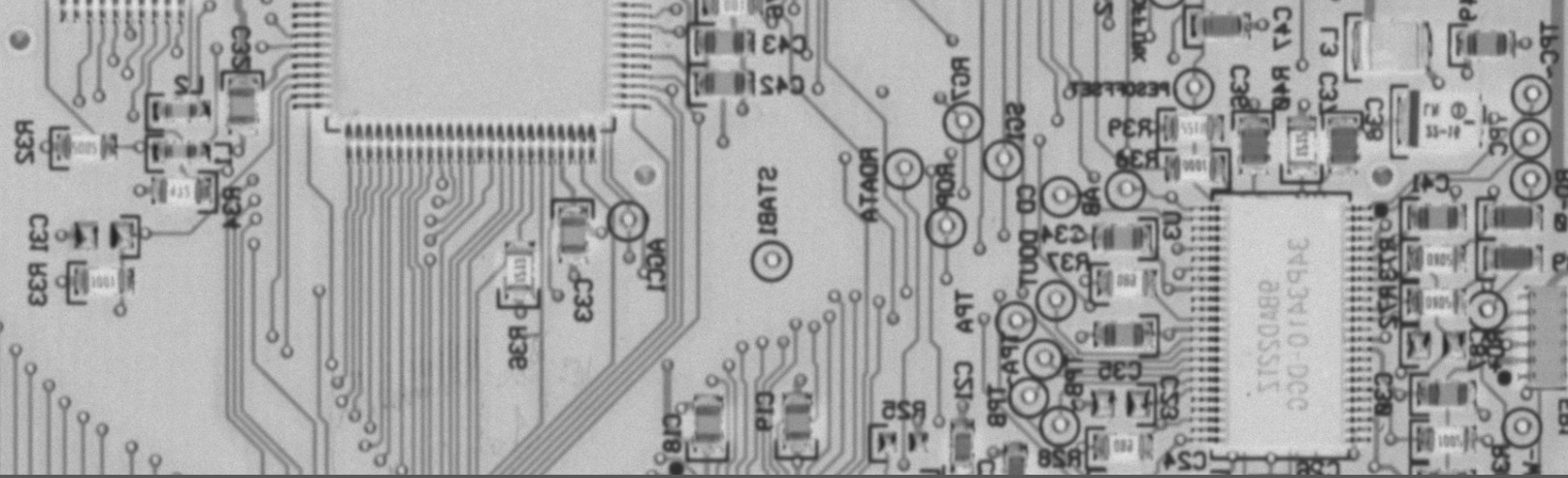
why this approach?



“One-size-fits-all”



Specialised to specific workloads

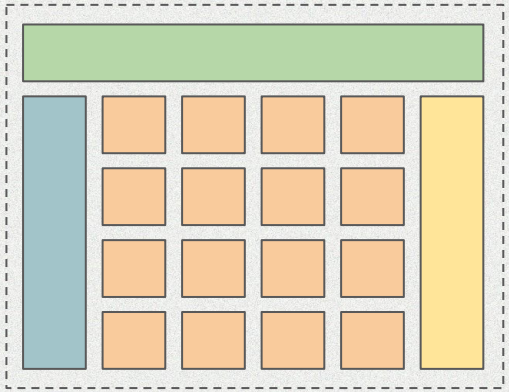


Accelerator Architecture

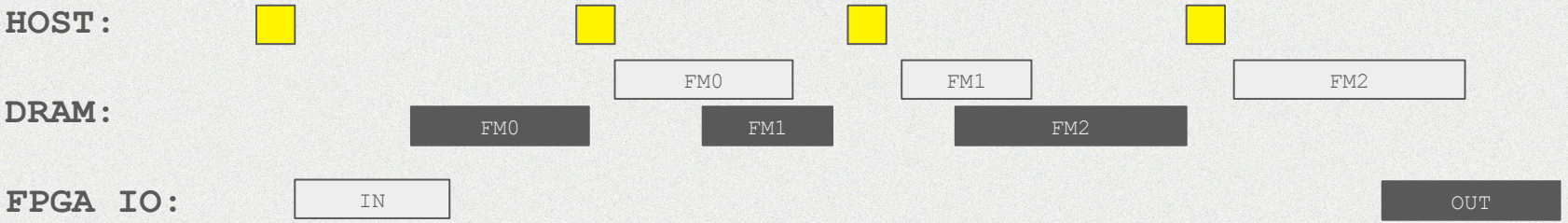
HERONIC
TECHNOLOGIES

ARCHITECTURE PRINCIPLES

Systolic Array Dataflow



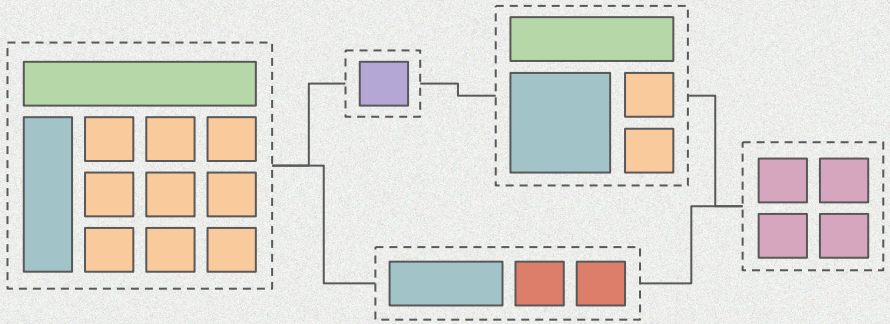
MAC Allocation	Homogeneous
Hardware Compilation	Overlay
Software Execution	Instructions
Parameter Storage	Off-Chip
Feature-map Storage	Off-Chip
Pipelining	Layer



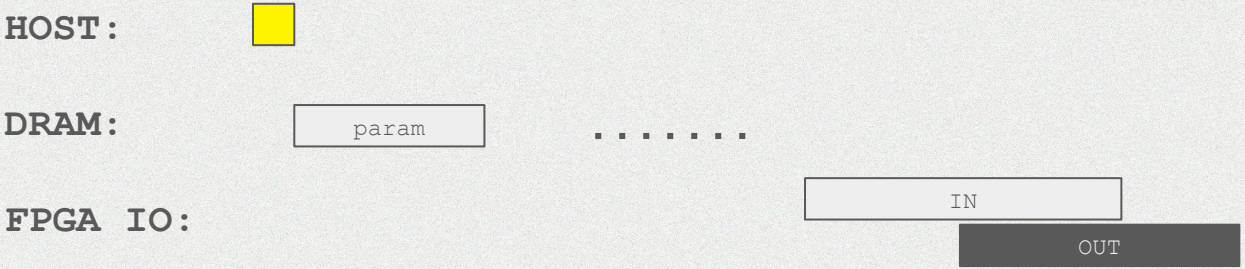
email: info@heronic.ai

ARCHITECTURE PRINCIPLES

Heronic (Streaming Dataflow)

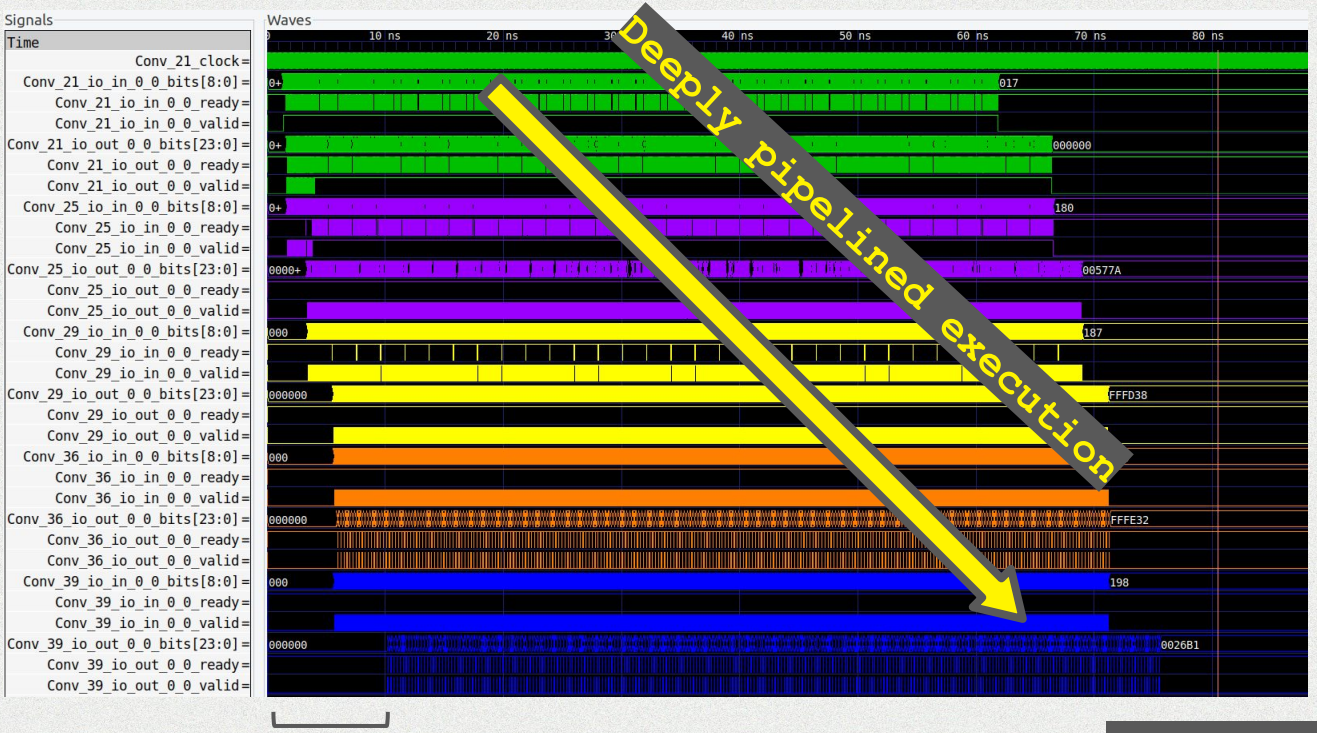


MAC Allocation	Heterogeneous
Hardware Compilation	Bespoke
Software Execution	Hostless
Parameter Storage	Hybrid
Feature-map Storage	On-Chip
Pipelining	Network

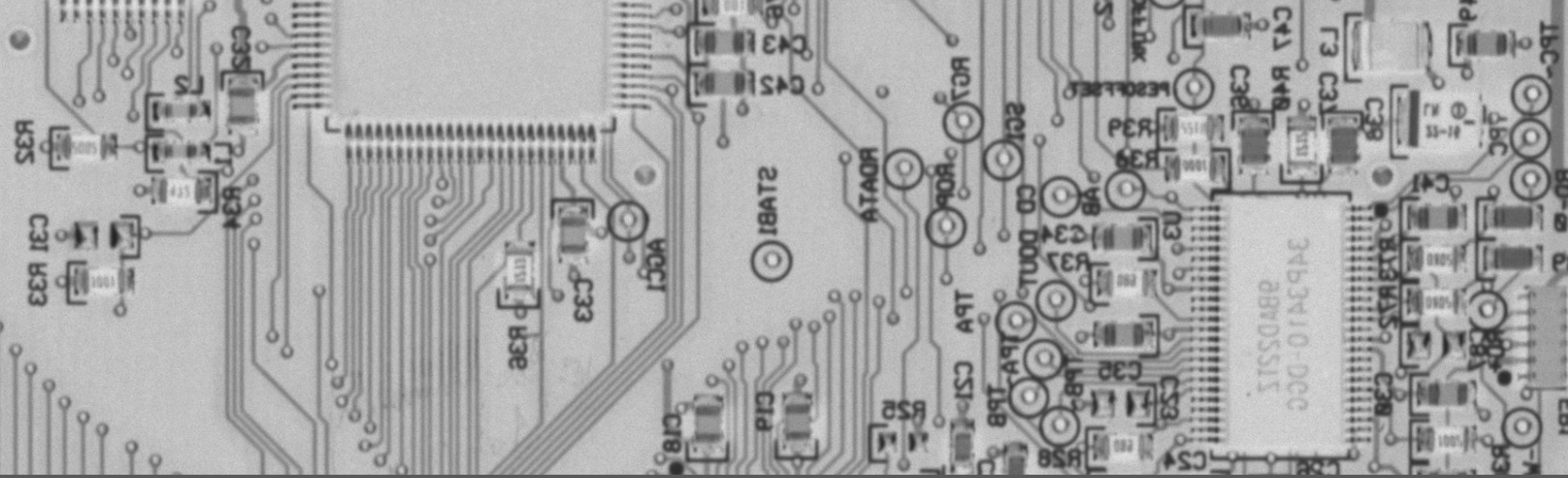


email: info@heronic.ai

EXAMPLE WAVEFORM



email: info@heronic.ai



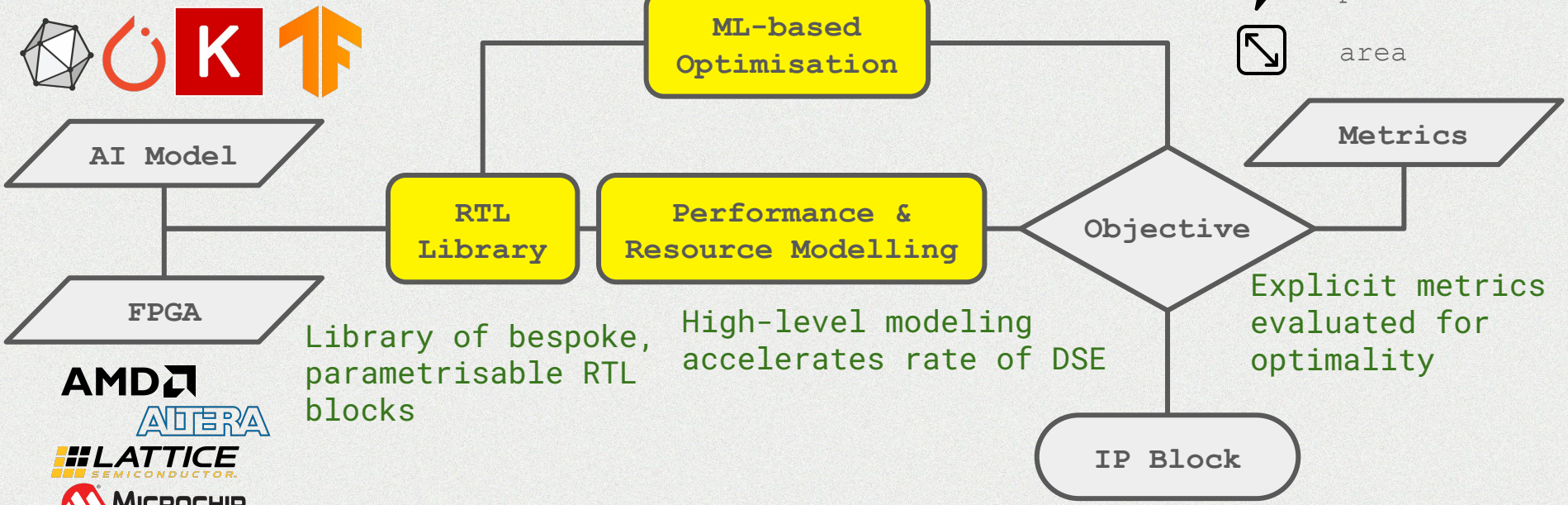
DSE Transforms

HERONIC
TECHNOLOGIES

AUTOMATION/MODELLING

Automating the design process

ML-driven DSE, finding optimal next design

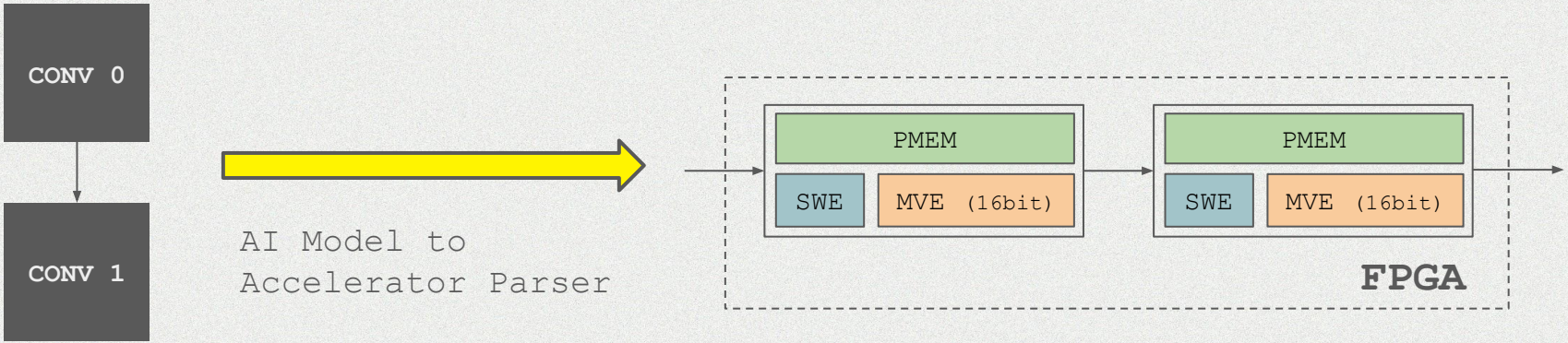


email: info@heronic.ai

BESPOKE HW BLOCKS

PMEM = Parameter Memory
SWE = Sliding Window Engine
MVE = Matrix-Vector Engine

What does our HW architecture look like?



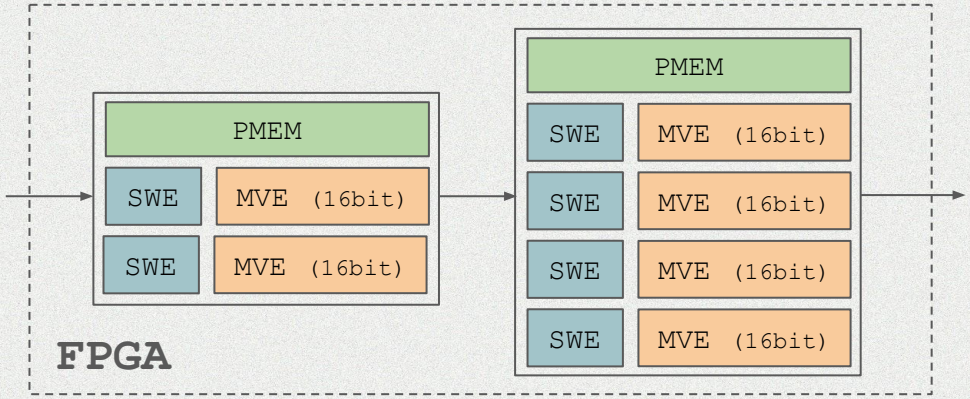
email: info@heronic.ai

LAYER LEVEL PARALLELISM

PMEM = Parameter Memory
SWE = Sliding Window Engine
MVE = Matrix-Vector Engine

Parallelism

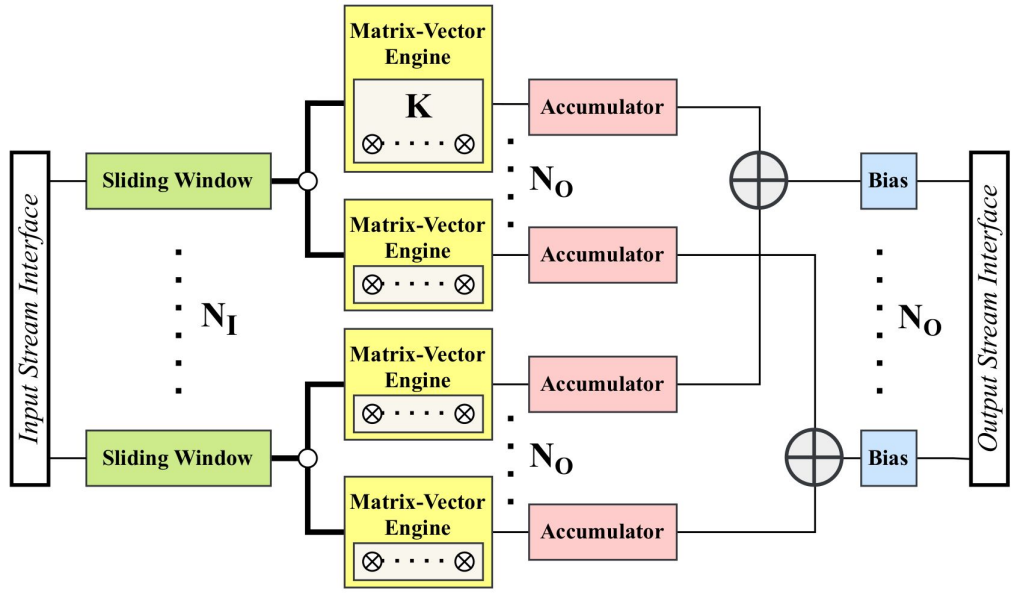
- Allocate resources to different layers
- Expose a resource-performance trade-off
- Leads to heterogeneous design



email: info@heronic.ai

LAYER LEVEL PARALLELISM

Convolution Hardware



NI = Input Parallelism
Integer < # Channels

NO = Output Parallelism
Integer < # Filters

K = Kernel Parallelism
Integer < # kernel size ^2

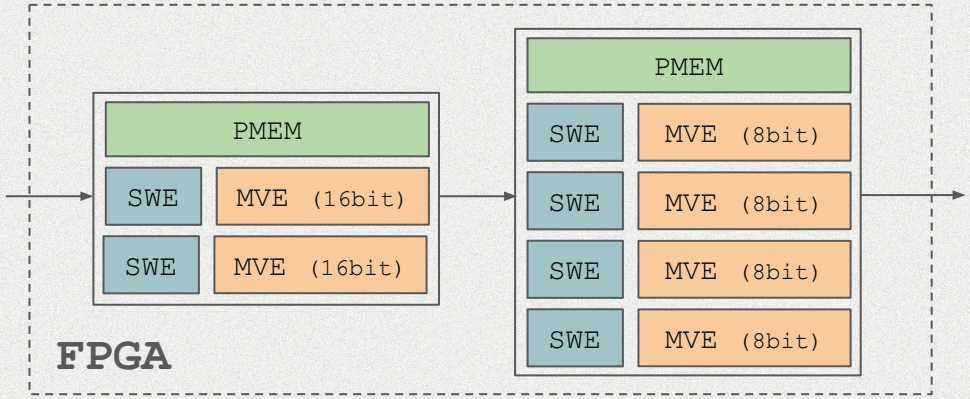
email: info@heronic.ai

QUANTISATION

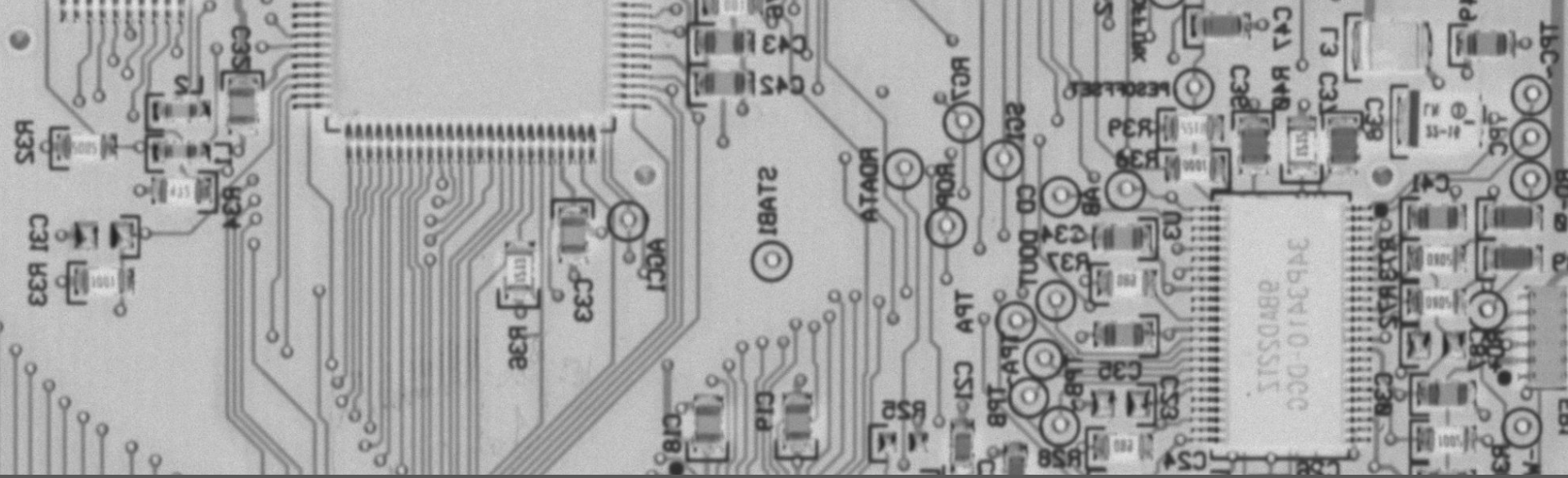
PMEM = Parameter Memory
SWE = Sliding Window Engine
MVE = Matrix-Vector Engine

Quantisation

- Change the wordlengths and quantisation methods per layer
- Accuracy vs resource trade-off
- Fine-grain control of model compression



email: info@heronic.ai



MLCommons Benchmark

HERONIC
TECHNOLOGIES

MLPERF TINY

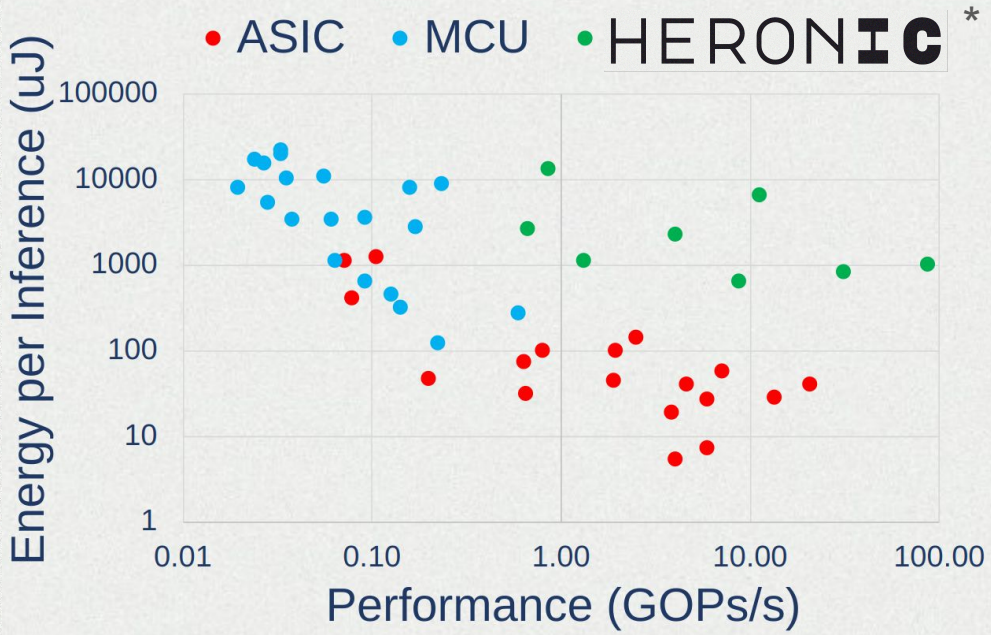
ML Commons

A benchmark run by MLCommons for AI running on embedded systems.

- Lowest latency for 3 benchmarks out of 24 submissions
- 8 accelerators designed in 2 months
- Fastest FPGA submission of all time

email: info@heronic.ai

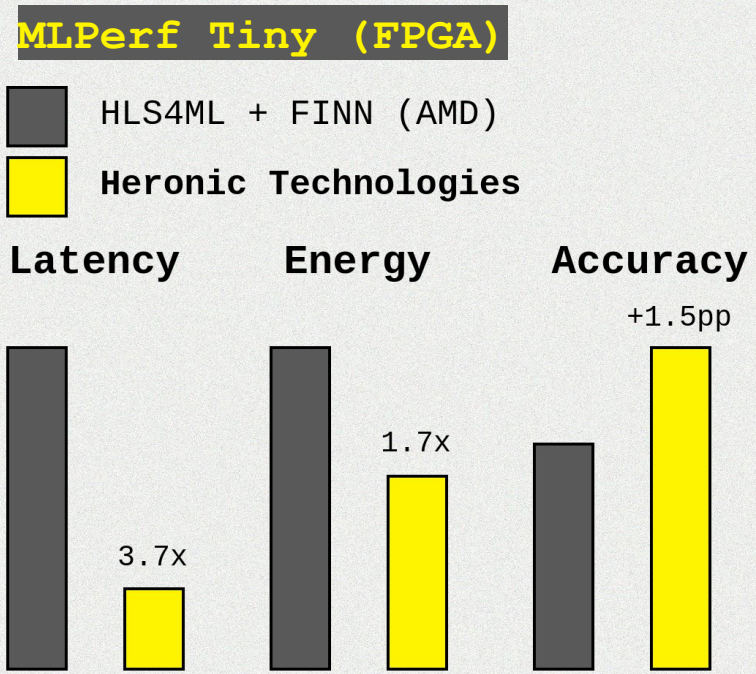
MLPERF TINY - RESULTS



*includes idle SoC power which is not used

email: info@heronic.ai

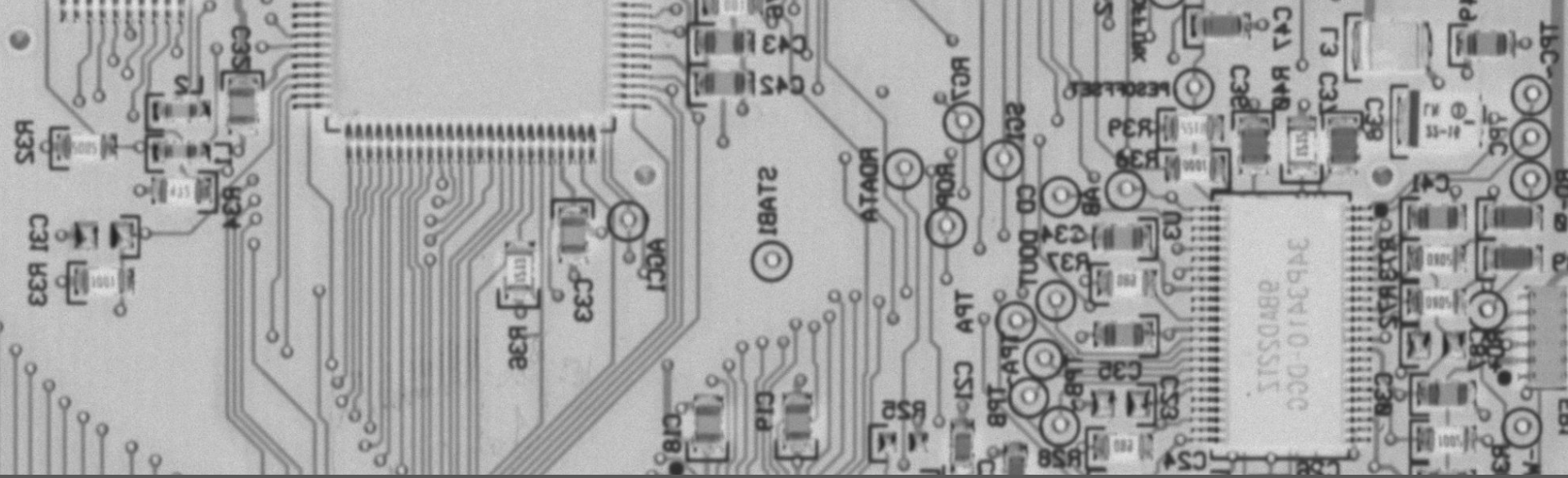
MLPERF TINY - RESULTS



- Better performance, energy and accuracy characteristics
- No modifications needed

	FINN	Heronic
Device	XC7Z020	XC7Z020
Quantisation	Custom (1-bit)	INT8
Model	Custom (CNV)	ResNet8

email: info@heronic.ai



Object Detection

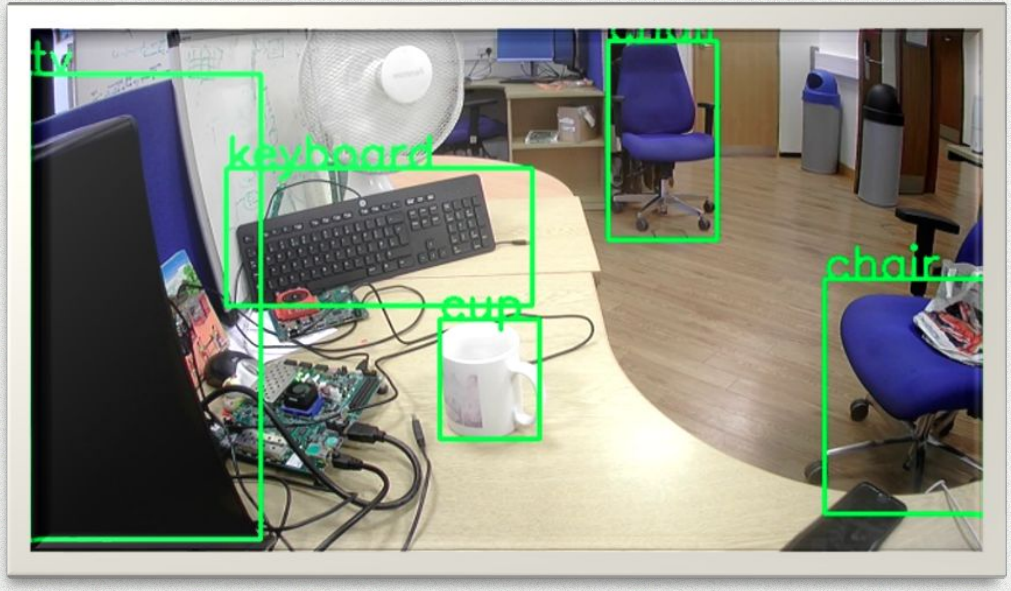
HERONIC
TECHNOLOGIES

OBJECT DETECTION

Targeting latest generation YOLO models

Challenges:

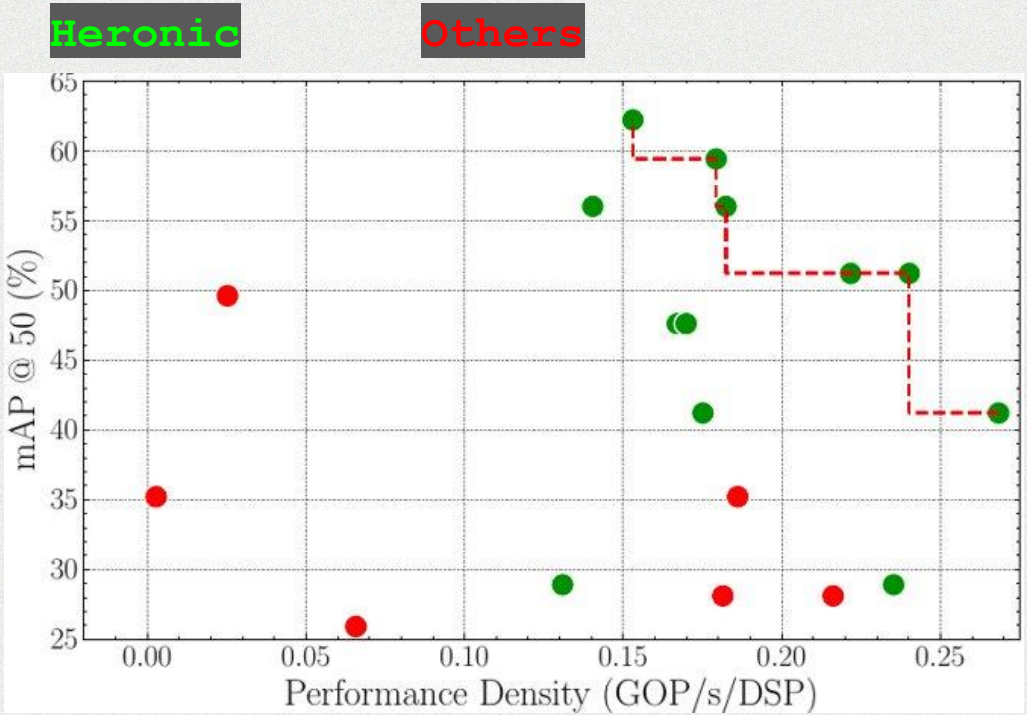
- Large number of parameters
- Heterogeneous architecture
- Deep skip connections



email: info@heronic.ai

OBJECT DETECTION

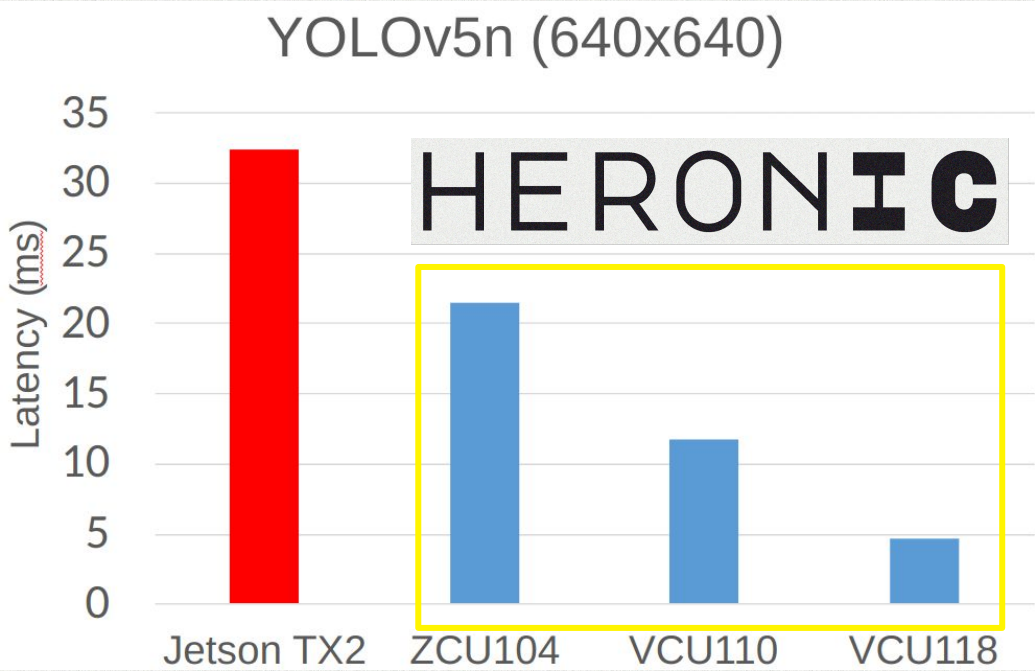
- DSE can find **pareto-optimal** points
- **Less constrained** on-chip memory
- **Higher** memory **bandwidth utilization**
- **Denser compute**
- Design automation means a **wider range** of models, and **SOTA support**



email: info@heronic.ai

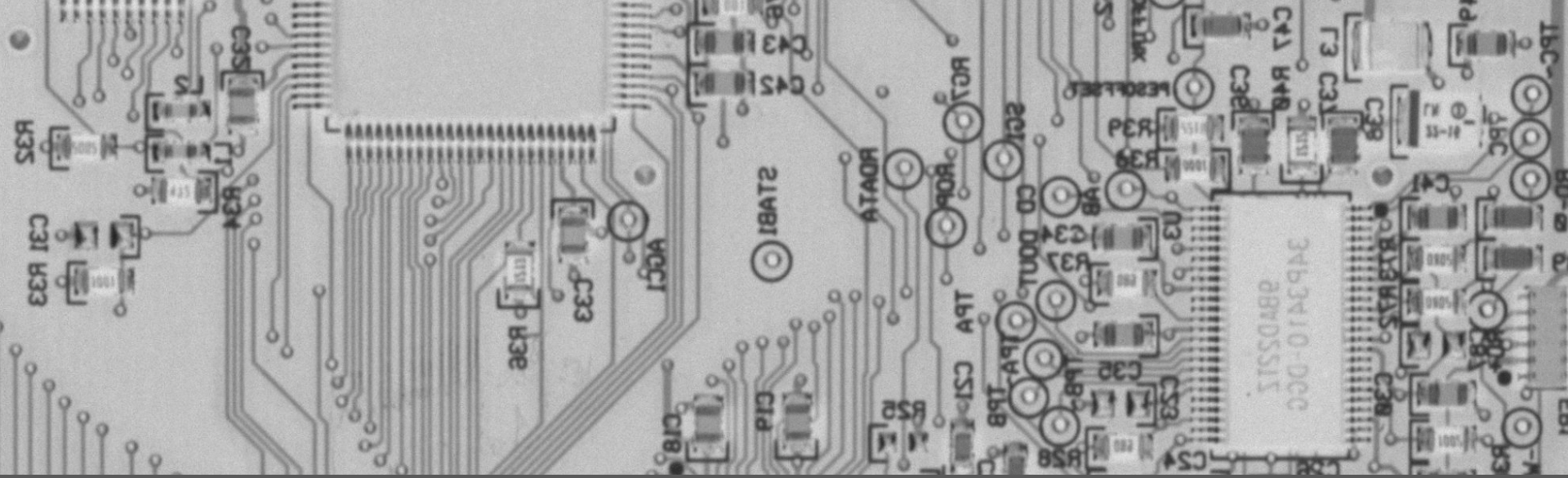
OBJECT DETECTION

Better than Nvidia (for certain cases)



- **Lower Latency** than GPU (up to 7x faster)
- **Similar energy** per inference (-10 mJ, +40 mJ)
- **Deeply pipelined**, so output arrives fast
- Not bottlenecked by **off-chip memory bandwidth**

email: info@heronic.ai



Thank you for listening

Alex Montgomerie
alex@heronic.ai

Chief Executive Officer

HERONIC TECHNOLOGIES