KLEPSYDRA

TECHNOLOGIES

THE Most Efficient Software for Edge Computing

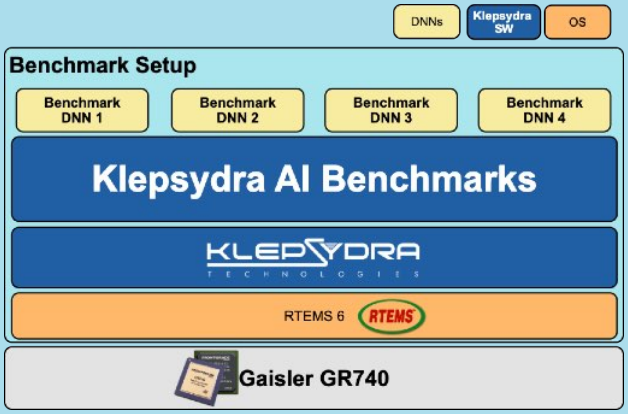
Unleash the Full Power of Your Space Computer

Klepsydra AI Performance Benchmark Report

Version: 19 **Target**: Gaisler GR740. **Date**: Dec. 2025

Technical setup

A Deep Neural Network (DNN) is trained using standard tools like TensorFlow or PyTorch and exported as an ONNX file. This file is then loaded into Klepsydra AI, running on the GR740.



Benchmark Deep Neural Networks

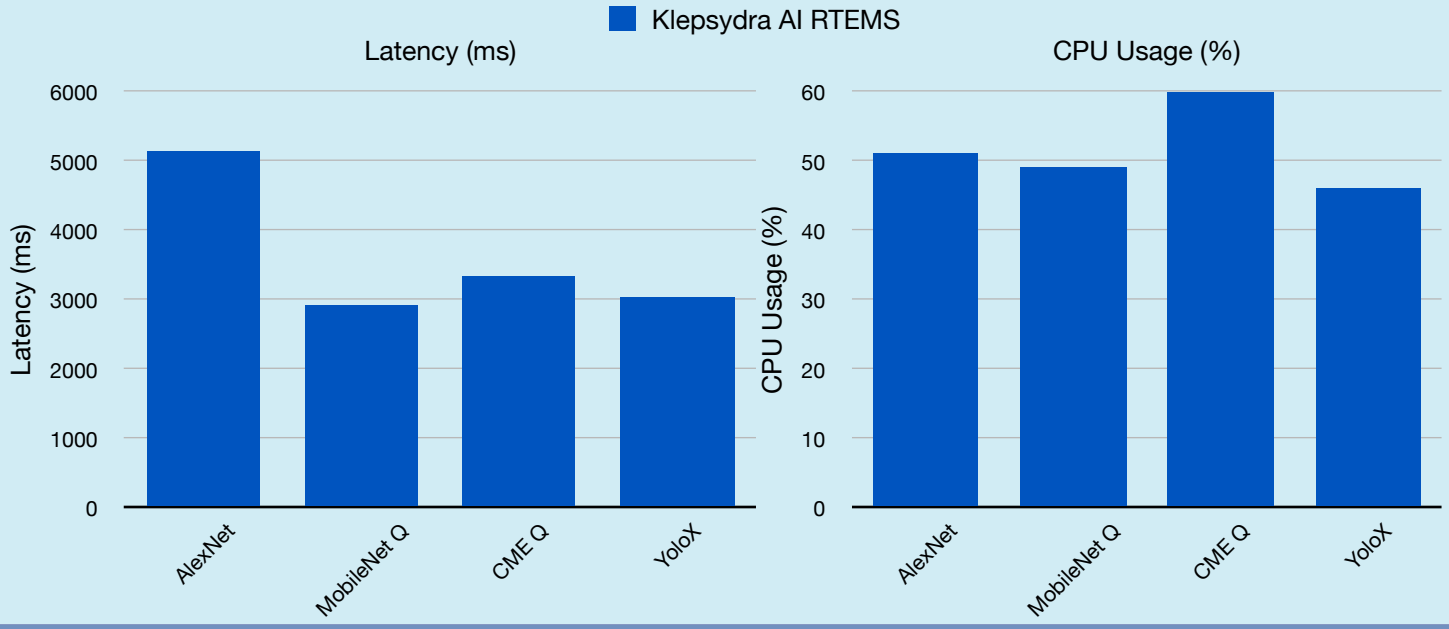
Several networks were tested as part of this campaign. These networks come from different sources as specified:

- AlexNet (open source)
- MobilNet Quantised
- ESA CME Quantised (ESA OBPMark-ML[1])
- YoloX / Ship Detection (ESA OBPMark-ML[1])

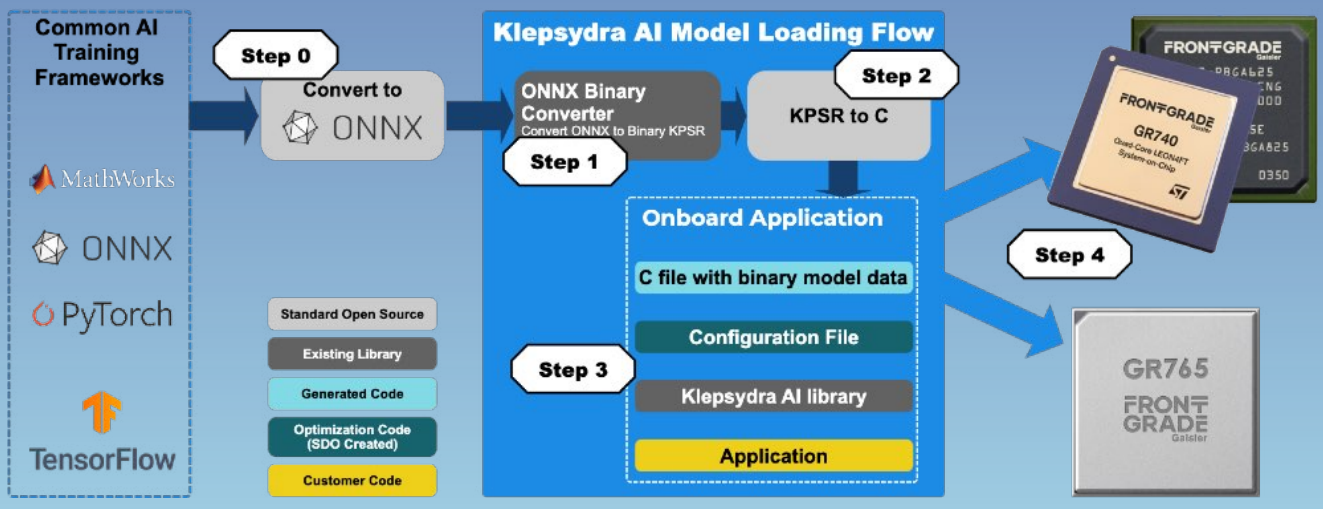
[1]: https://zenodo.org/records/5638577

The performance results are shown for latency, i.e., the time required to execute the AI algorithm for a given input data, and CPU consumption, i.e., how much CPU is used for executing the AI algorithm.

TensorFlow Lite does not run on GR740 - Klepsydra AI is currently the only AI inference framework capable of of running AI algorithms on the GR740.



Easy to use architecture for integrating AI into space systems

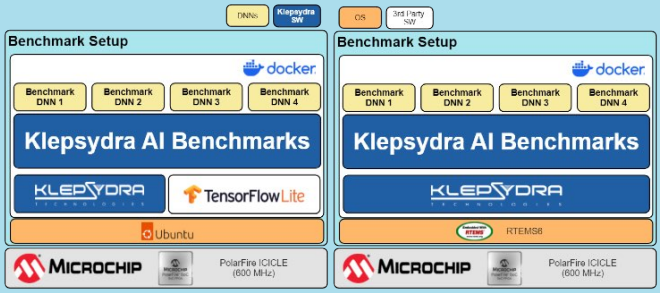


Klepsydra AI Performance Benchmark Report

Version: 18 **Target:** PolarFire RISC-V. **Date:** Dec. 2024

Technical setup

The benchmark application runs in a Docker container on Ubuntu 22.04 on a PolarFire Icicle RISC-V, with the image including Klepsydra AI and TensorFlow Lite 2.4.4 (compiled with NEON extensions). For RTEMS*, the AI benchmarks are deployed directly on the target binary.



Benchmark Deep Neural Networks

Several networks were tested as part of this campaign. These networks come from different sources as specified:

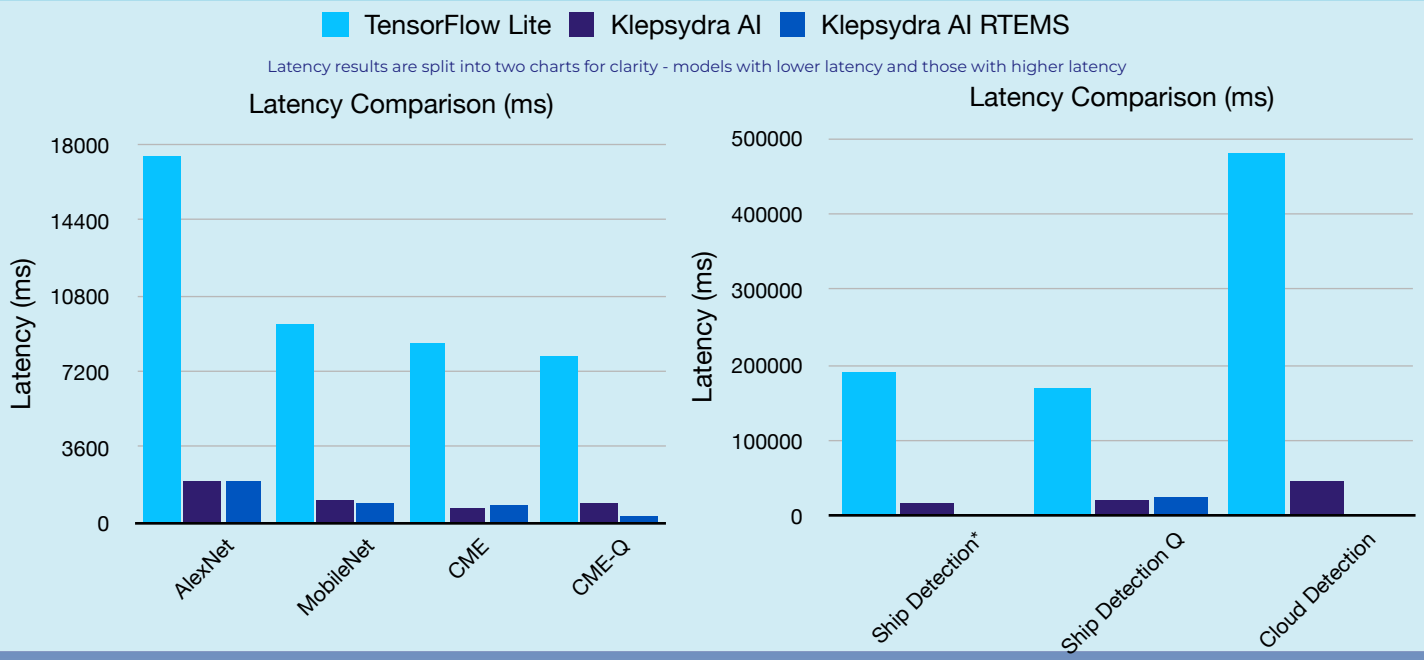
- AlexNet (open source)
- ESA Coronal Mass Ejection Detection (CME [1])
- ESA CME Quantised (ESA OBPMark-ML[1])
- YoloX / Ship Detection* (ESA OBPMark-ML[1])
- YoloX / Ship Detection Quantised (ESA OBPMark-ML[1])
- Cloud Detection (ESA OBPMark-ML[1])

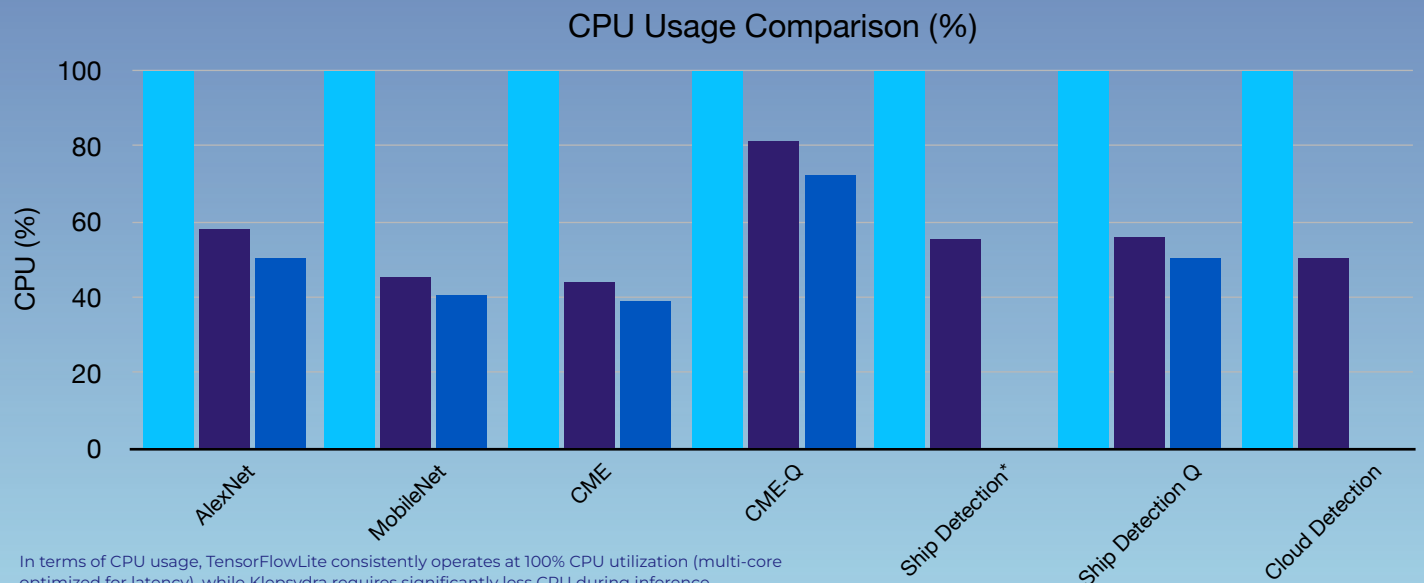
[1]: https://zenodo.org/records/5638577

* Benchmarks for RTEMS are partially complete, with a full benchmark report expected in a new document release in Q2 2025.

The performance results are shown for latency, i.e., the time required to execute the AI algorithm for a given input data, and CPU consumption, i.e., how much CPU is used for executing the AI algorithm.

The results show that Klepsydra AI outperforms TensorFlowLite in terms of latency and CPU consumption for the RISC-V architecture.





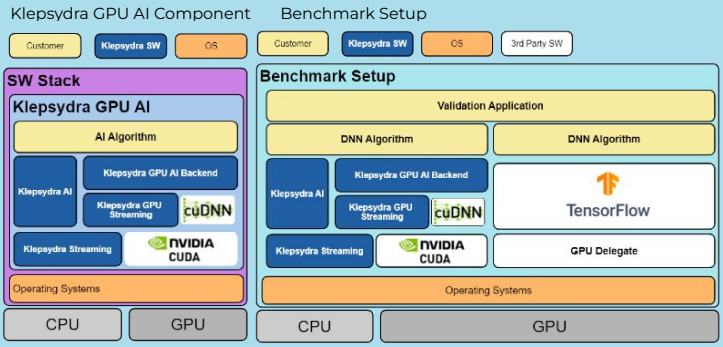
In terms of CPU usage, TensorFlowLite consistently operates at 100% CPU utilization (multi-core optimized for latency), while Klepsydra requires significantly less CPU during inference.

Klepsydra AI Performance Benchmarks Report

Version: 18 **Target:** NVIDIA TX2i. **Date:** Dec. 2024

Technical setup

The GPU benchmark was carried out with Klepsydra AI for GPU on a TX2i.
The total binary size was less than 5Mb including Klepsydra libraries and the validation software.



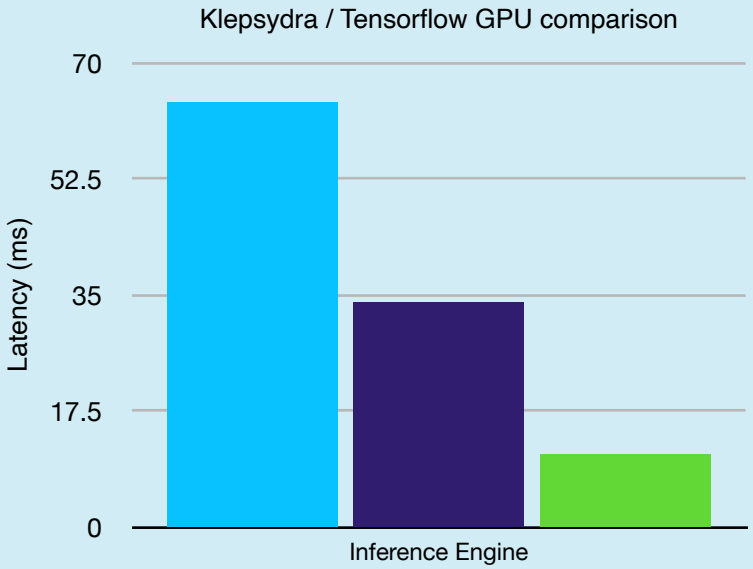
Benchmark Deep Neural Networks

The performance of Klepsydra GPU AI has been validated using a Deep Neural Network to perform signal-noise-optimization, proivded by the European Space Agency.

The validation itself was carried out validating several preloaded datasets from previous collections and execute the example repeatedly for each of these datasets.

The validation software allows to inject different AI inference engines. Thus Klepsydra AI and TensorFlow can be compared in terms of performance and precision.

The results show that Klepsydra AI outperforms TensorFlowLite in terms of latency:



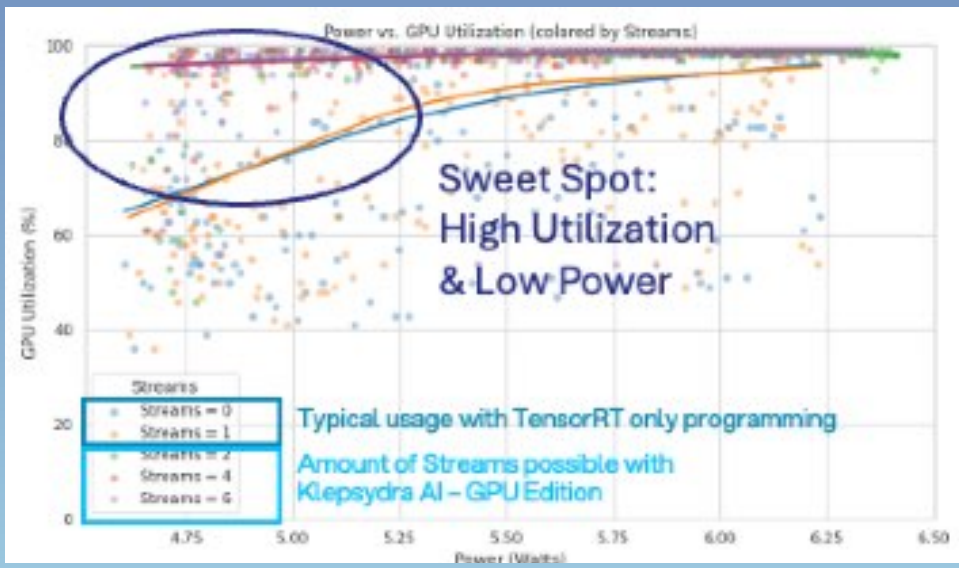
	NVIDIA Tools	Klepsydra GPU Connector
Inference engine	TensortRT	GPU Connector (CuDNN-based) & TensorRT
GPU Occupancy	Medium	High
Cyber-security	No	Yes
Safety Certification	No	Yes (For the CPU part)
Secure OS Support	No	VxWorks, DDC-i (For the CPU part)

Klepsydra can orchestrate in an intelligent and efficient manner the use of parallel streams in the GPU.

Different tasks can be run on the same GPU with the support of Klepsydra:

- Individual kernels
- Cudagraphs (groups of kernels)
- TensorRT

Klepsydra patent-pending algorithm for GPU orchestration relies on our patented eventloop for high efficient, zero memory copy orchestration.

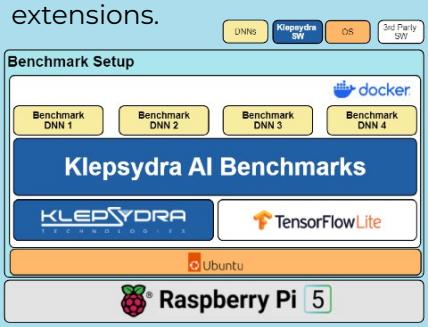


Klepsydra AI Performance Benchmarks Report

Version: 18 **Target**: Raspberry Pi5. **Date**: Dec. 2024

Technical setup

The benchmark application is run on a docker container on top of the Ubuntu 22.04 on the Raspberry Pi5 4-core CPU. The docker image contains both Klepsydra AI as well as TensorFlowLite 2.4.4 compiled with NEON extensions.



Benchmark Deep Neural Networks

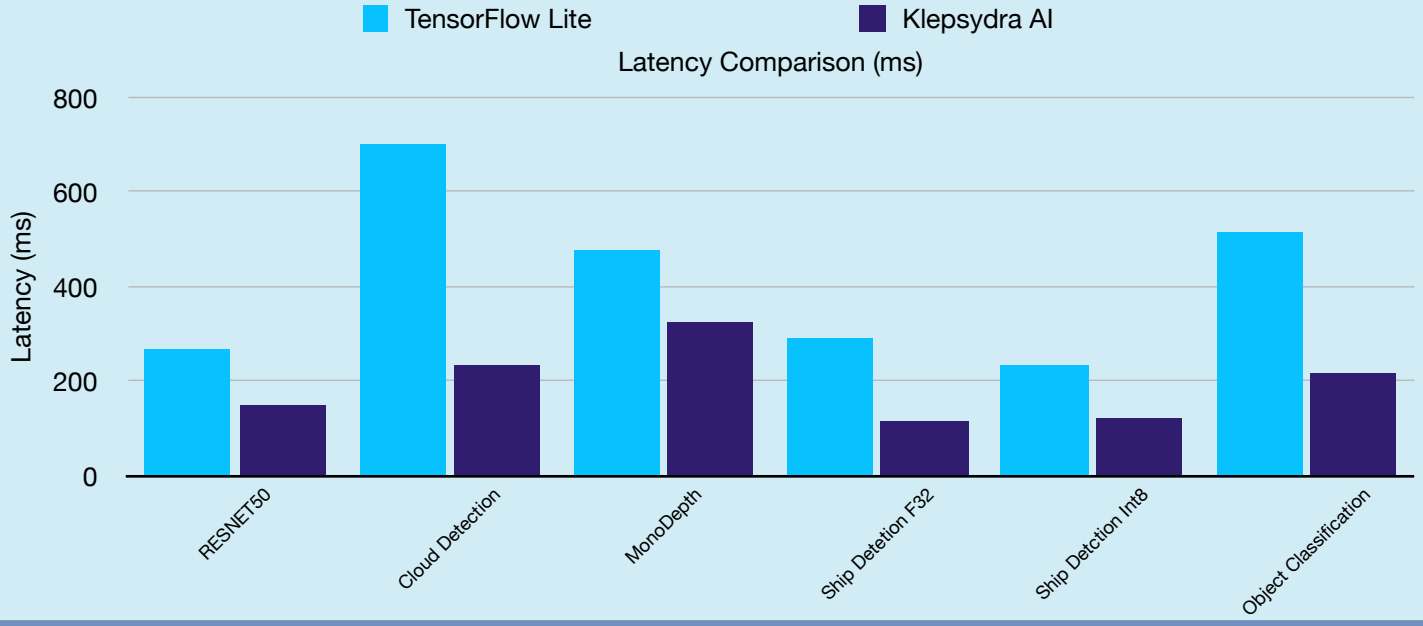
Several networks were tested as part of this campaign. These networks come from different sources as specified:

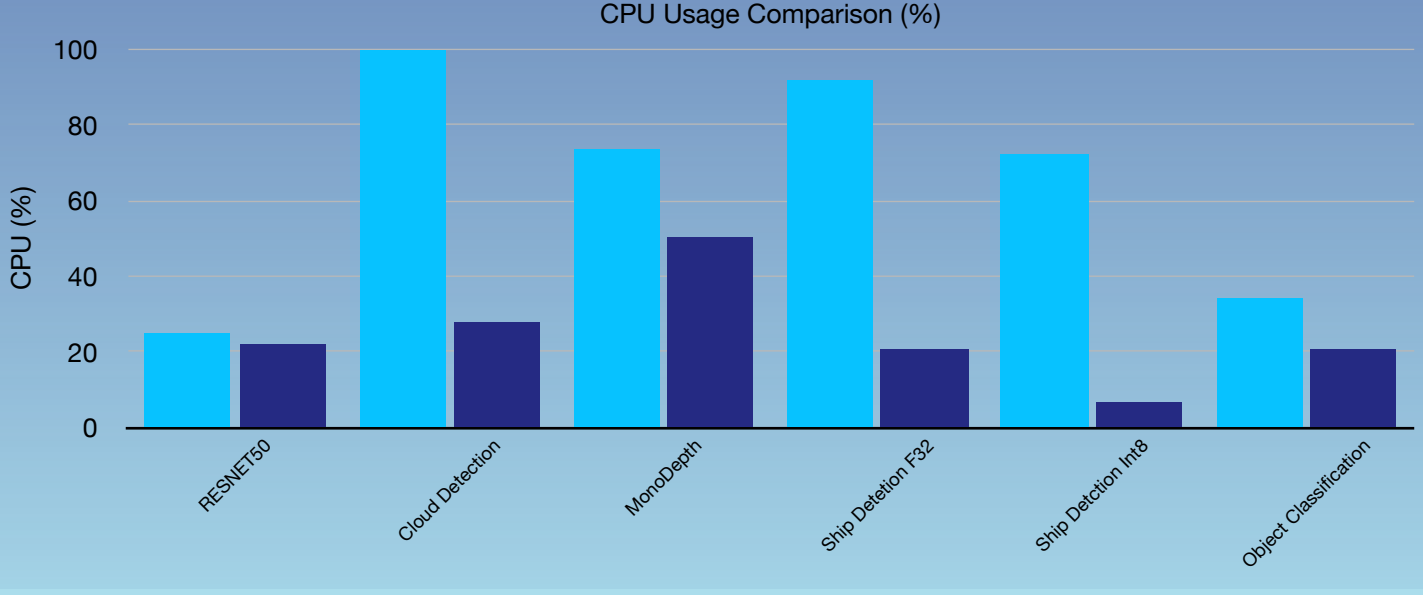
- RESNET50 (MLAB Project)
- Cloud Detection Quantised (ESA OBPMark-ML[1])
- Monodepth (open source)
- YoloX / Ship Detection (ESA OBPMark-ML[1])
- YoloX / Ship Detection Quantised (ESA OBPMark-ML[1])
- Object Classification (D'Amico paper[2])

[1]: https://zenodo.org/records/5638577
[2]: https://arxiv.org/abs/2309.11645

The performance results are shown for latency, i.e., the time required to execute the AI algorithm for a given input data, and CPU consumption, i.e., how much CPU is used for executing the AI algorithm.

The results show that Klepsydra AI outperforms TensorFlowLite in terms of latency and CPU consumption for the Raspberry Pi5.



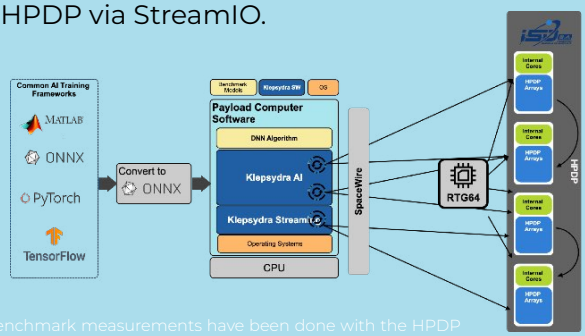


Klepsydra AI Performance Benchmark Report

Version: 19 **Target:** HPDP* **Date:** Mar. 2025

Technical setup

A Deep Neural Network (DNN) is trained using standard tools like TensorFlow or PyTorch and exported as an ONNX file. This file is then loaded into Klepsydra AI, running on the payload processor. The subset layers of the DNN that will be offloaded to the HPDP are sent to the HPDP via SpaceWire so that the FNC can load them into the XPP (code and configuration). Alternatively, this subset of layers can be loaded into the RTG4 RISC-V processor for execution in the HPDP via StreamIO.



* All benchmark measurements have been done with the HPDP HW Simulator due to limited access to the actual processor

Benchmark Deep Neural Networks

Several networks were tested as part of this campaign. These networks come from different sources as specified:

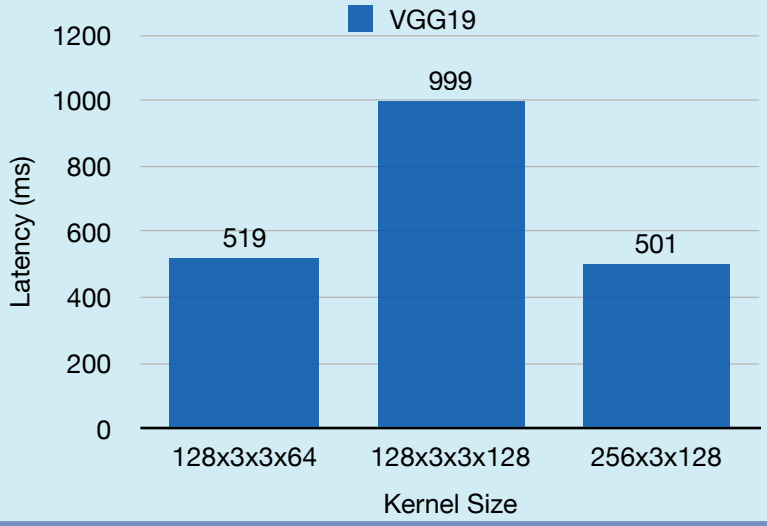
- ResNet50 Quantised (MLAB Project)
- MobileNetV2 Quantised
- YoloX / Ship Detection Quantised (ESA OBPMark-ML[1])
- VGG19 Quantised (MLAB Project)

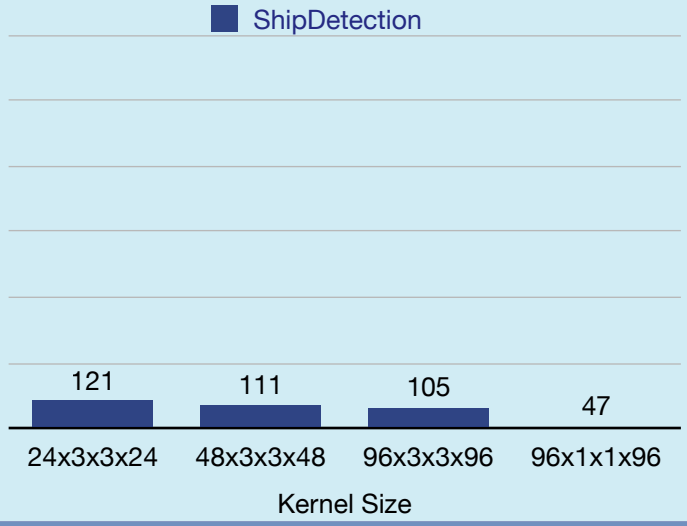
New Capabilities

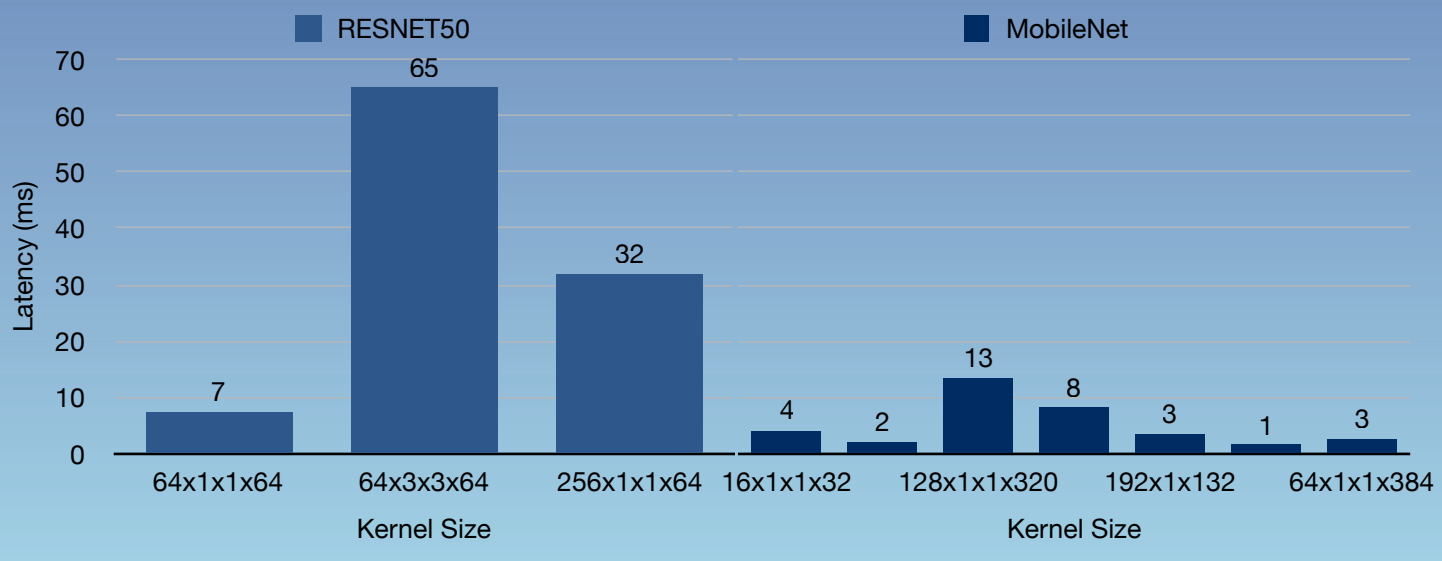
- Enable execution of AI on HPDP, a radiation hardened processor.
- Easy to use: Execute AI in 2-3 lines of code.
- No requirement for development of NML code for the HPDP: users can simply provide their existing AI models.
- Interoperability solution with Machine Learning (ML) frameworks: user can train AI models in any framework and provide the AI model file to payload computer + HPDP for execution without any development effort.

No other inference engine runs on HPDP

Klepsydra AI is the only AI inference framework capable of of running AI algorithms on the HPDP.





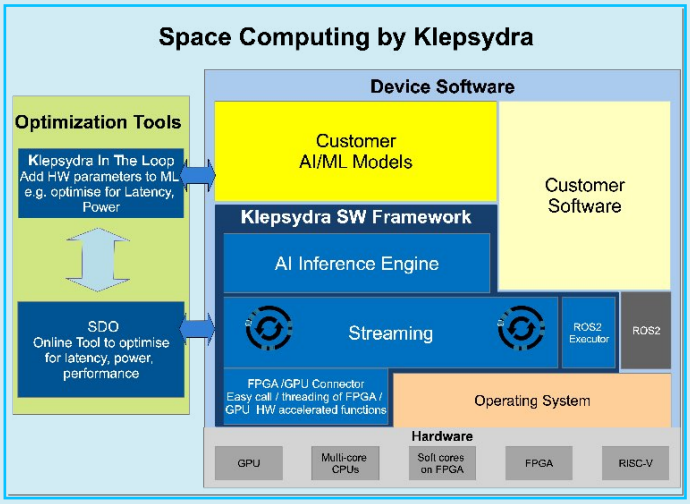


Klepsydra AI.
High performance AI for space

Klepsydra AI is a high-performance deep neural network engine designed for edge computing. It allows customers to deploy both existing and newly trained models on space computers, just like with typical edge AI solutions. Klepsydra AI delivers four key advantages:

- Boost data processing performance
- Reduce power consumption
- Supports major AI formats and software
- Simple adoption and integration

Klepsydra Framework



Klepsydra Streaming
Boost data processing for general and processor-intensive algorithms.

Klepsydra AI
A high-performance deep neural network engine to deploy AI/ML models across various processors, including CPU-only.

GPU/FPGA Connector
Maximize data throughput and GPU utilization with high parallelization. Easily integrate FPGA acceleration to enhance performance.

ROS2 Executor
Processes up to 10x more data while cutting CPU consumption by 50%.

Klepsydra Streaming Distribution Optimizer
A configurable framework to optimize throughput for CPU, GPU, or FPGA-based algorithms.

Klepsydra in-the-Loop
Integrates hardware performance into model training for improved efficiency and accuracy.

How Klepsydra Enhances Your Space AI Capabilities?

Our optimized Software Framework lets you:
- Process 10x more data with the same processor and AI model
- Cut power usage by 50%, with the same AI model
- Shorten deployment and development time

Develop YOUR software and AI on our platform
➡ We offer the acceleration framework,
YOU build the system

Higher accuracy and reliability: Klepsydra AI offers greater stability, predictability, and determinism than other edge solutions.

Integration

Wide compatibility: supports most common AI formats and software; deployable on various edge devices like ARM CORTEX-A, RaspberryPi, Intel NUC, TX2i, VITIS AI FPGA, etc.

Easy adoption: the intuitive API simplifies integration, and its unique visual autotuning interface lets users easily optimize models for specific devices.

Cost Reduction

Avoid extra costs: Customers process more data with less energy, avoiding the need for more powerful solutions to run edge AI.

Optimized costs: For new hardware, customers can fine-tune solutions to handle real data with reduced hardware size and lower energy consumption, leading to cost savings.

Klepsydra AI employs a high-performance 2D parallelization model, enabling high-precision models on the edge with just 3 lines of code.

- **The Application API** is a straightforward asynchronous API using the predict-callback pattern.
- **The Dynamic Backend API** enables various hardware accelerators and follows the strategy pattern.

KLEPSYDRA

TECHNOLOGIES

Fast, secure and safe AI software for Space and Defence systems

Software is available at
https://downloads.klepsydra.com/
Or contact
sales@klepsydra.com

Available in three formats:

- Docker images
- Debian packages
- Tarballs

A trial license covers:

- A 90-day license
- 10 hours of technical support
- A 1-hour technical workshop

Contact

Klepsydra Technologies AG
6300 Zug, Switzerland

www.klepsydra.com
sales@klepsydra.com

+Klepsydra Technologies
wa.me/41782493720