KLEPSYDRA

TECHNOLOGIES

THE Most Efficient Software for Edge Computing

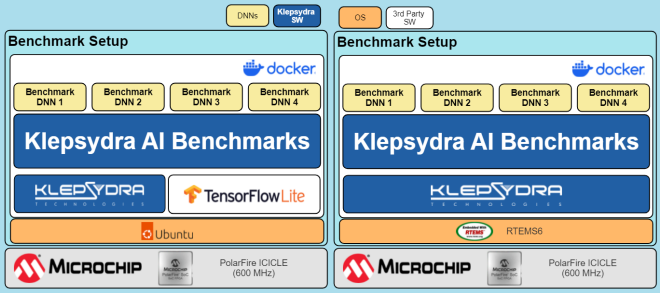
Unleash the Full Power of Your Edge Processor

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:

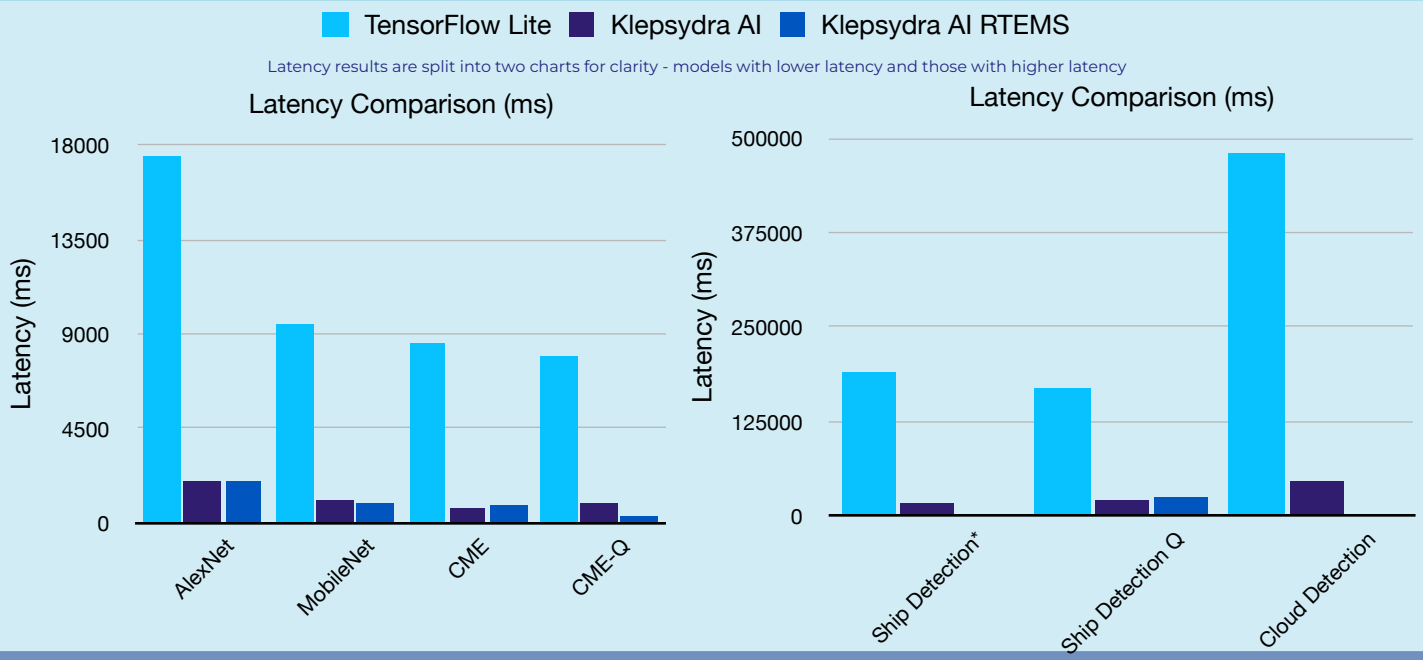
- AlexNet (open source)
- ESA Coronal Mass Ejection Detection (CME [1])
- ESA CME Quantised (ESA OBPMark-ML[1])
- YoloX / Ship Detection* (ESA OBPMark-ML[1])
- YoloX / Ship Detection Quantised (ESA OBPMark-ML[1])
- Cloud Detection (ESA OBPMark-ML[1])

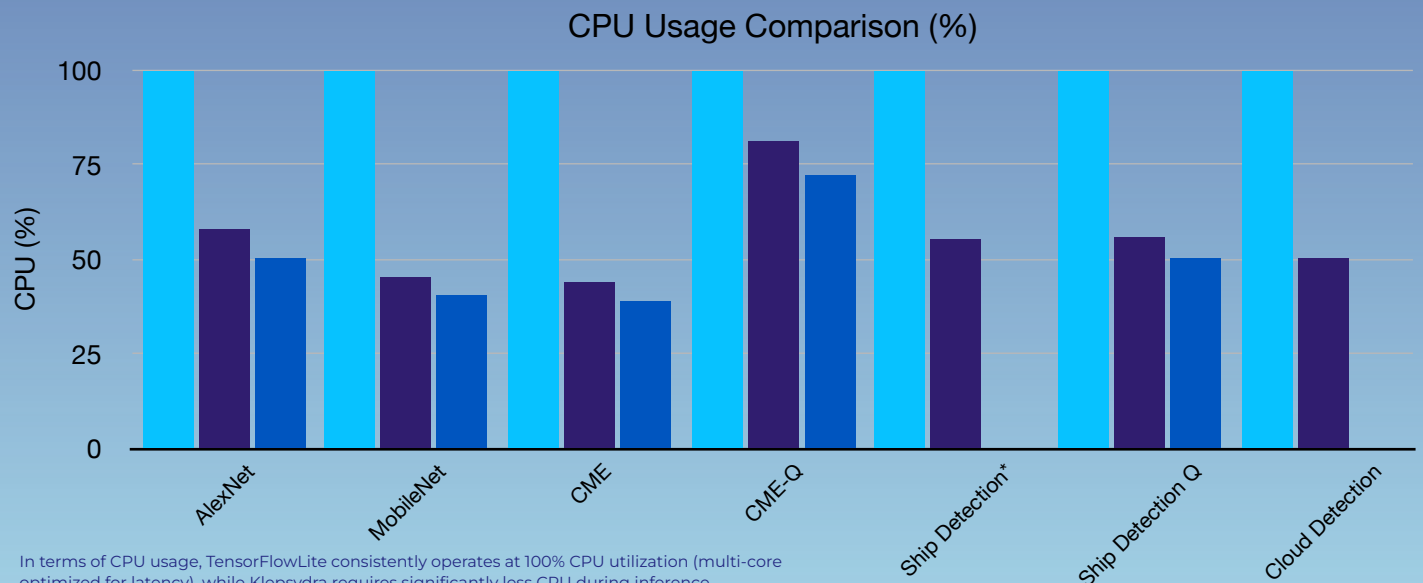
[1]: https://zenodo.org/records/5638577
[2]: https://arxiv.org/abs/2309.11645

* 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.





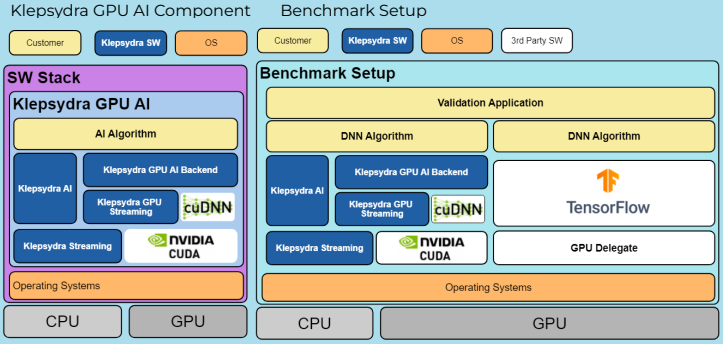
In terms of CPU usage, TensorFlowLite consistently operates at 100% CPU utilization (multi-core optimized for latency), while Klepsydra requires significantly less CPU during inference.

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

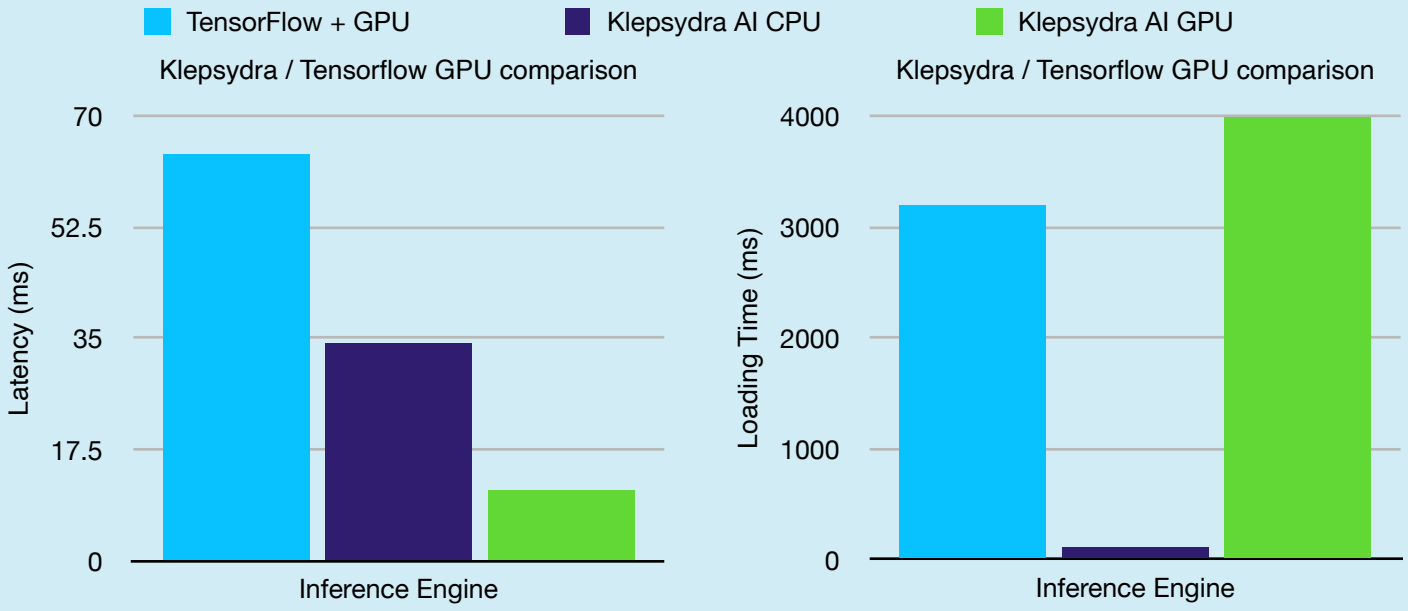
The performance of Klepsydra GPU AI has been validated using a Deep Neural Network to perform signal-noise-optimization, proivded by the European Space Agency.

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 performance results are shown for latency, i.e., the time required to execute the AI algorithm for a given input data, and loading time, i.e., how long it takes to load the AI algorithm.

The results show that Klepsydra AI outperforms TensorFlowLite in terms of latency, regarding loading time, see comment below.



Conclusion related to the usage of GPUs - Observation on the load times:

While the GPU version of Klepsydra AI offers enhanced computational capabilities, the loading time for the GPU inference engine is 30 to 40 times longer than that of the CPU inference engine. This is driven by the fact that always the Capture API needs to be executed to run the full network. **As the validation application was built to execute short intermittent burst of data, the Capture API is called every time with a new burst resulting in long load times.**
This performance characteristic highlights that the choice of processing architecture in edge applications depends heavily on the specific use case:

- For short, intermittent bursts of data or stop-and-start AI applications, CPU-only processors with Klepsydra AI are recommended as the impact of the load time over processing speed is significantly higher.
- For applications involving frequent model switching, CPU-only processors with Klepsydra AI are preferable.
- For long-duration tasks with high data rates and large data volumes, GPUs running Klepsydra GPU AI are the optimal choice.

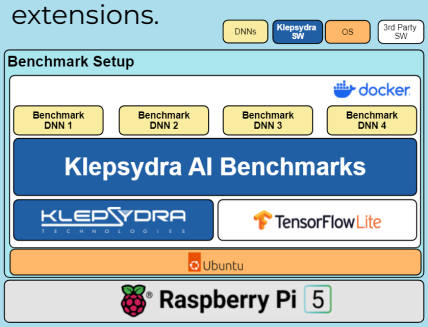
These insights underline the adaptability of Klepsydra AI to diverse computational needs related to edge AI.

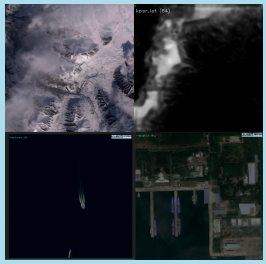
Klepsydra AI Performance Benchmarks Report

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

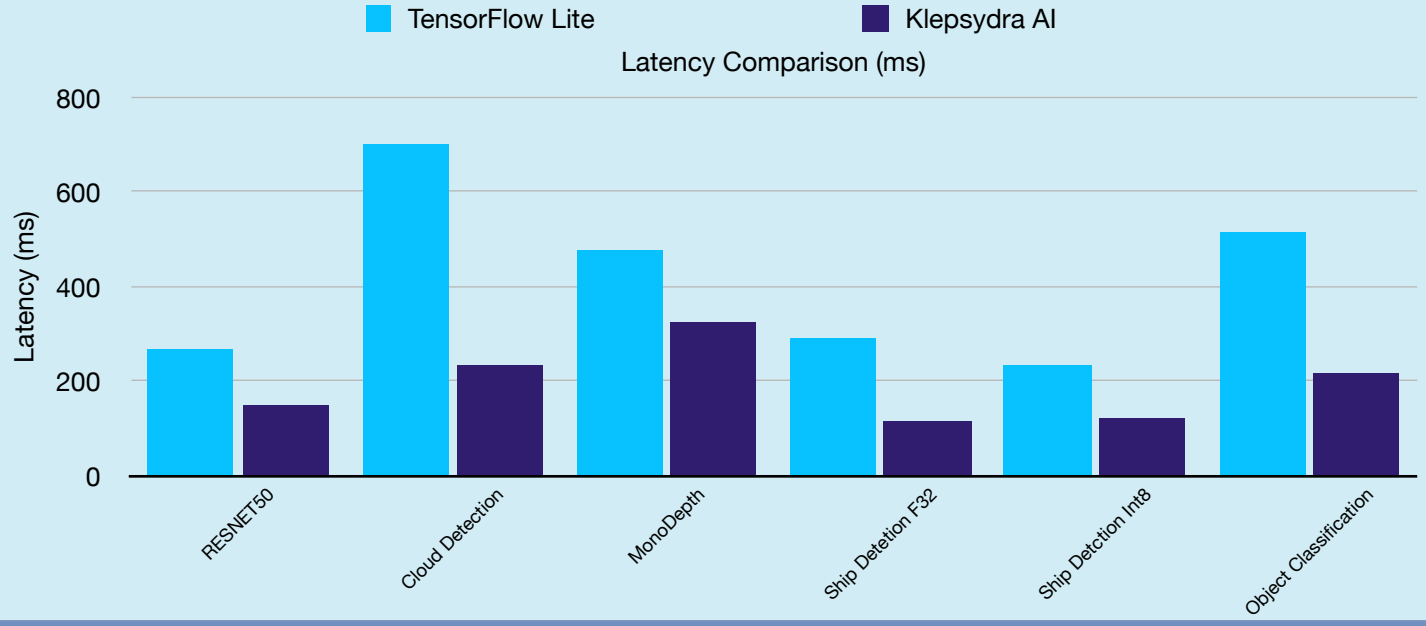
Several networks were tested as part of this campaign. These networks come from different sources as specified:

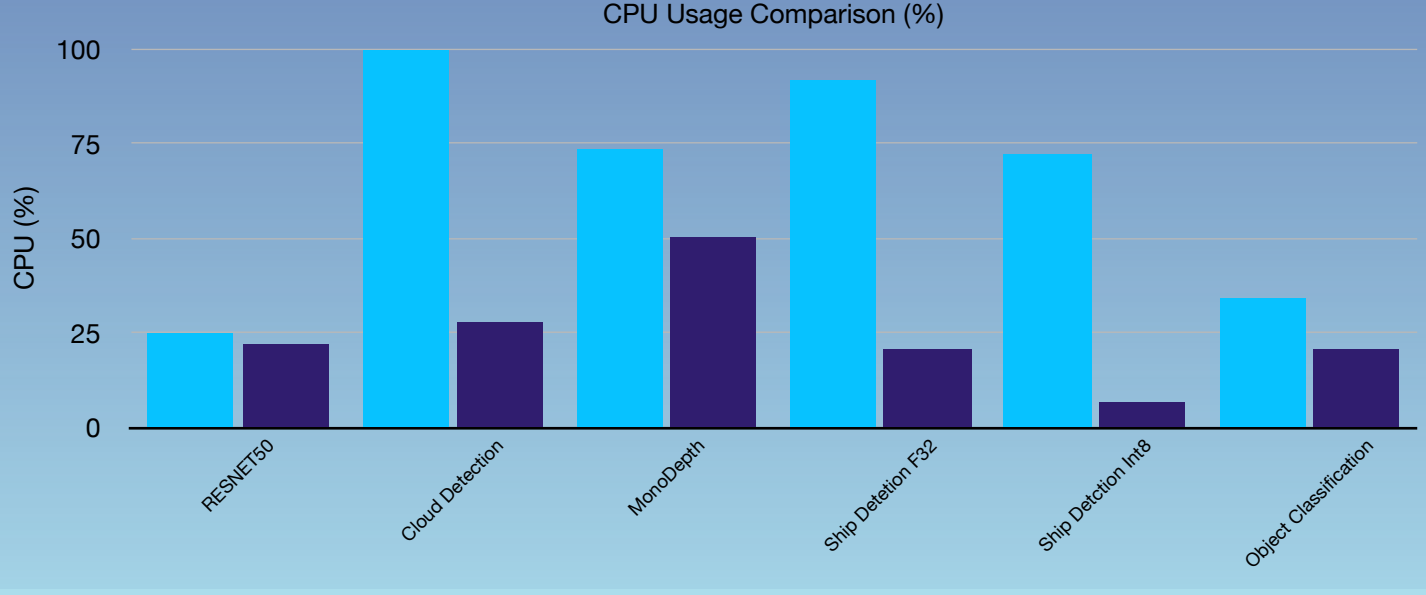
- RESNET50 (open source)
- Cloud Detection Quantised (ESA OBPMark-ML[1])
- Monodepth (open source)
- YoloX / Ship Detection (ESA OBPMark-ML[1])
- YoloX / Ship Detection Quantised (ESA OBPMark-ML[1])
- Object Classification (D'Amico paper[2])

[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.



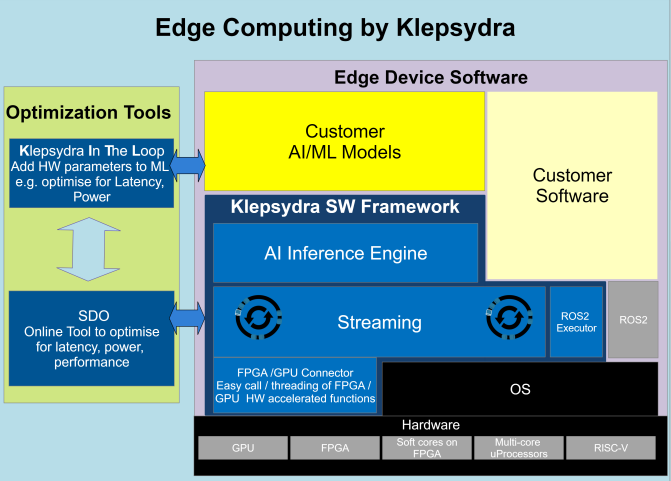


Klepsydra AI.
High performance AI for the edge

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 at the edge, 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 Edge 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

Bring Intelligence to a Broader Range of Embedded Systems

Contact

Klepsydra Technologies AG
6300 Zug, Switzerland

www.klepsydra.com
sales@klepsydra.com

+Klepsydra Technologies
wa.me/41782493720