Klepsydra AI – Product Brief

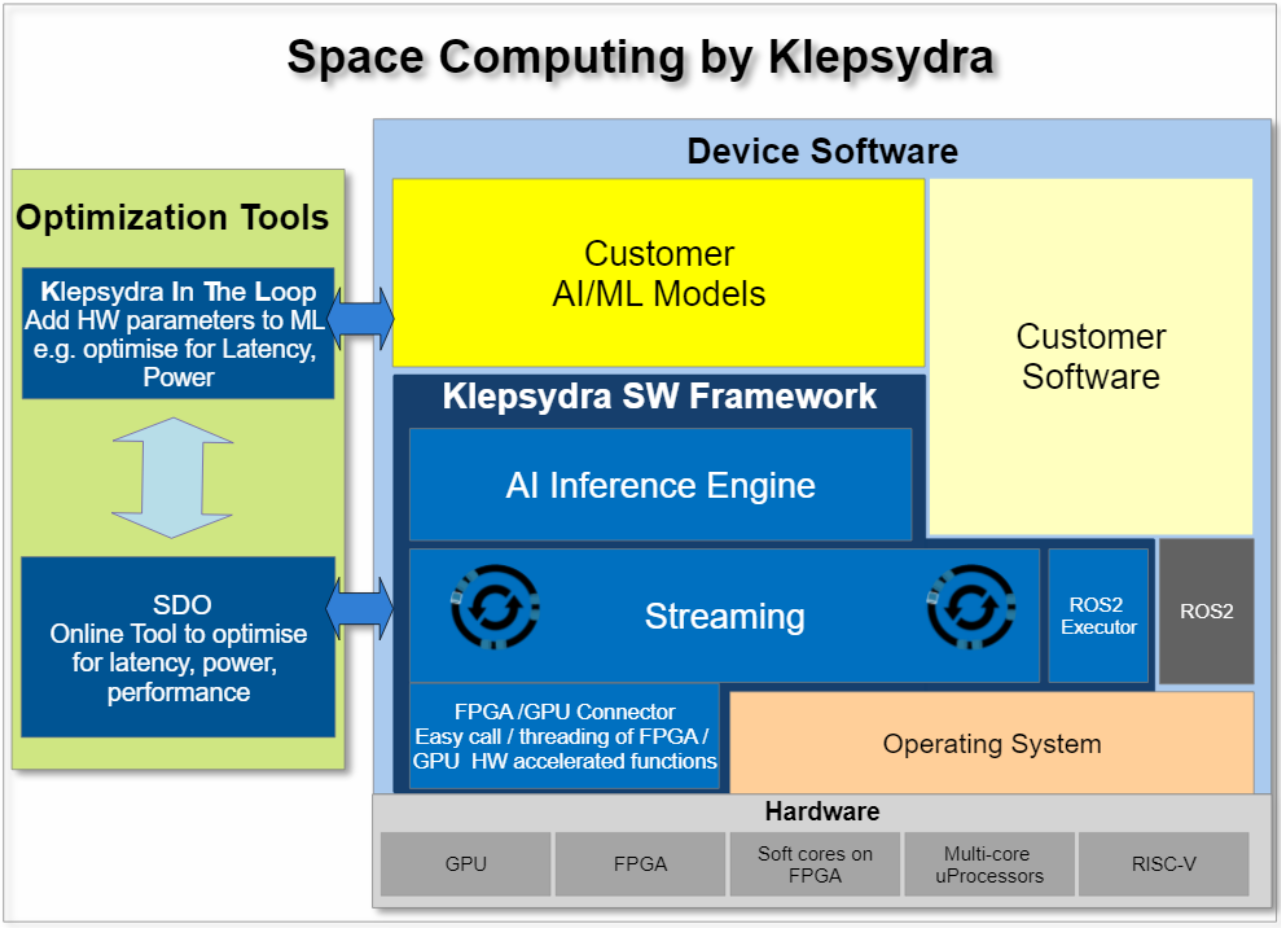
April 28, 25

Introduction

Klepsydra's highly efficient data processing and AI Inference framework is a key building block to enable Intelligent Space System.

Enabling the necessary data processing and AI inference capabilities on many different space computers and processor, the Klepsydra SW framework provides a common SW framework to enable the data processing performance needed for intelligent space systems, regardless of the underlying onboard processor, therefore reducing development costs. And the powerful AI inference engine allows to bring AI to all types of space vehicles and subsystems on different processors with a unified API. And by separating AI inference from the AI/ML models, while getting the necessary performance, it provides a common flexible architecture for different AI/ML based applications in space.

Klepsydra software is a patented, modular and high performance framework that allows running AI inference on standard space computers without the need for dedicated hardware accelerators:



As the framework can also utilise hardware accelerators when they are available it extends the hardware choices that can be used for intelligent Space System and therefore allows to implement Intelligence on a larger range of space systems, from low cost systems using low cost processors to high-end systems with dedicated accelerator hardware.

Programming on the Klepsydra framework is hardware independent and thus allowing entities to have a more generic approach towards software defined function and running AI in an space systems:

Product Description

Klepsydra Streaming
Boost data processing at the edge

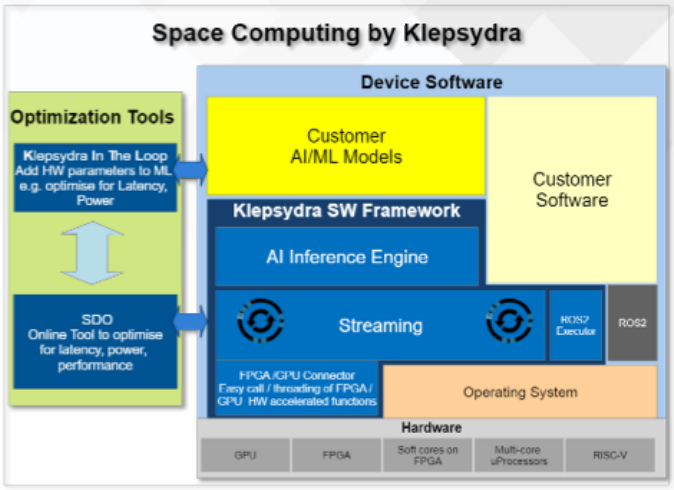
Klepsydra AI – Artificial Intelligence
High performance inference, including CPU-only.

GPU / FPGA Connector
Easy integration of GPU& FPGA HW acceleration

ROS2 Executor plugin
10 x more data with up to 50% reduction in CPU usage

Klepsydra Streaming Distribution Optimiser
Configure framework for maximum data throughput and algorithms running on CPU, GPU, or FPGA

Klepsydra in-the-Loop
Include hardware performance into model training



Klepsydra Streaming

At the core of the framework is Klepsydra Streaming, an advanced SW library that optimizes data streaming using a combination of pipelining and parallelization. Its modular, LEGO-like structure enables developers to build complex workflows with elastic optimization for latency, CPU usage, and throughput. Supporting sensor multiplexing and event loop processing, Klepsydra Streaming operates asynchronously, eliminating the need for scheduling and reducing memory overhead for greater efficiency.

Using lock-free algorithms, Klepsydra Streaming avoids power-intensive context switching, instead relying on atomic operations like Compare-and-Swap (CAS) to manage shared resources. This approach ensures higher throughput, lower power consumption, and consistent CPU performance—critical for real-time applications.

Particularly effective in tasks like matrix multiplications, Klepsydra Streaming delivers seamless, high-performance solutions for resource-constrained edge environments.

Klepsydra AI

Klepsydra AI is a powerful AI inference engine that delivers groundbreaking performance for real-time edge-device applications. It abandons traditional vectorization methods like OpenMP in favor of a pipeline-based approach, where each thread processes a specific part of the algorithm, such as matrix multiplications. This methodology allows Klepsydra AI to execute AI models on low-power onboard computers with up to 4x the efficiency of market leaders like TensorFlowLite while consuming up to 50% less energy.

The engine has been successfully validated through the European Space Agency's (ESA) KATESU project, where it achieved outstanding results on Teledyne e2v's LS1046 and Xilinx ZedBoard computers. Klepsydra AI enables even computationally constrained devices to handle complex AI workloads, making it a transformative tool for applications requiring real-time AI processing.

Klepsydra GPU/FPGA Connector

The Klepsydra GPU/FPGA Connector extends the frameworks capabilities to HW accelerators, offering a hybrid optimization strategy that splits workloads between GPUs or FPGAs and CPUs. By solving GPU occupancy problems and leveraging a C++/CUDA backend, the connector ensures compatibility with most NVIDIA boards while significantly enhancing algorithm performance. The FPGA connector is still in the early TRL stage, but it has been tested and benchmarked on the Xilinx KRIA KV260 and Ultrascale+ platforms, where it demonstrated exceptional results in AI inference for navigation tasks. Klepsydra Streaming orchestrates the process, ensuring efficient communication between CPUs and HW accelerators. We will further work on the FPGA connector to prepare it for the market.

Klepsydra SDO (Streaming Distribution Optimizer)

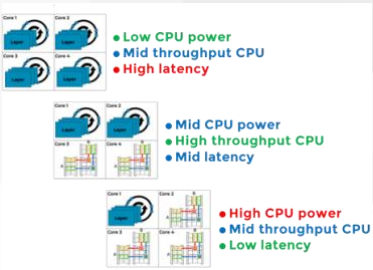
Klepsydra SDO is a tool that complements KSF's on-device capabilities by optimizing latency, throughput, and CPU usage. Developers can upload logs from dry runs of their applications to SDO, which then runs an optimizer to determine the ideal configuration for their specific use case. The tool generates a configuration file that can be implemented on the target HW, ensuring the best trade-offs for performance. With SDO, developers can optimize for single or dual combinations of latency, throughput, and CPU usage, simplifying application calibration.

Technology

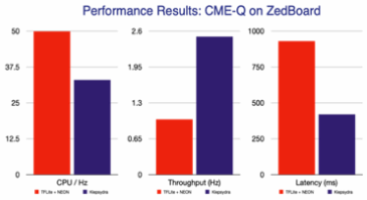
The acceleration technology of the Klepsydra SW framework is based on combining a sophisticated two-dimensional parallelization system with lock-free programming. The result is a flexible and efficient inference engine that can be optimized for three performance criteria in respect to other engines:

- **Latency:** achieving up to 4 times latency reduction
- **Throughput:** achieving up to 10 times increase in the data processing rate
- **Power consumption:** achieving up to 50% reduction

Careful configuration of the parallel threads allows allocating different memory and CPU resources to each element of the AI algorithm, which can help optimize the performance of the AI inference for latency, throughput or CPU utilization.



As part of the Klepsydra solution, we have developed the autotuning tool SDO which tests the target DNN with different configurations and finds the best configuration for each optimization constraint - low latency, high throughput, low CPU and low memory requirements.



Klepsydra has carried out several projects with the European Space Agency (ESA) proving that the technology yields excellent result for Space onboard computers running Space-related AI algorithms.

Value Proposition

- **High Throughput:** Klepsydra enables edge devices to process up to 10 times more data than traditional approaches, allowing for efficient and scalable solutions in data-intensive applications. This capability dramatically increases computational capacity without requiring HW upgrades.
- **Reduced Latency:** By reducing latency by up to 50%, Klepsydra ensures faster response times for individual tasks. This low-latency performance is critical for real-time applications where rapid decision-making is essential.
- **Power Efficiency:** Klepsydra reduces processor power consumption by up to 50% supporting power management on space system.
- **Cost Efficiency**: Klepsydra enables customers to maximize the performance of legacy HW, processing more data with less energy and avoiding costly HW upgrades. Its advanced deep neural network engine allows AI/ML models to run efficiently even on CPU-only systems, eliminating the reliance on GPUs for non-complex applications. By leveraging onboard CPUs for AI, Klepsydra reduces costs and energy consumption while broadening the accessibility of AI across different space systems.
- **HW-Agnostic**: Klepsydra facilitates AI deployment across diverse edge devices, from quad-core CPUs and SoCs to HW accelerators and multichip systems, offering broad compatibility with popular HW and SW configurations. By decoupling AI inference from AI/ML models and supporting standard formats like ONNX, Klepsydra eliminates vendor lock-in, enabling seamless portability across HW environments.
- **Deterministic Performance**: Klepsydra ensures consistent and predictable performance by maintaining flat CPU usage and throughput throughout execution. This deterministic behavior is essential for applications that demand stable, long-term performance under varying workloads.
- **Ease of Use**: Klepsydra simplifies algorithm integration and migration with its intuitive interface and efficient API, enabling quick adoption and accelerated deployment, even on legacy systems. The visual autotuning graphical interface further streamlines performance optimization for AI/ML models, making fine-tuning for specific HW effortless.
- **Enhanced Resource Management**: Klepsydra gives developers complete control over CPU resources, enabling them to balance throughput, latency, and CPU consumption based on specific application requirements. Additionally, Klepsydra has a small file-system footprint, which minimizes storage demands and allows it to operate efficiently even on resource-constrained systems.
- **Low Maintenance**: By separating AI/ML models from the execution framework, Klepsydra enables over-the-air updates with minimal risk and low bandwidth requirements. Model files as small as 4 MB streamline updates without altering the underlying SW stack. Additionally, open-source components promote community-driven innovation and ensure compatibility with diverse HW and SW ecosystems.

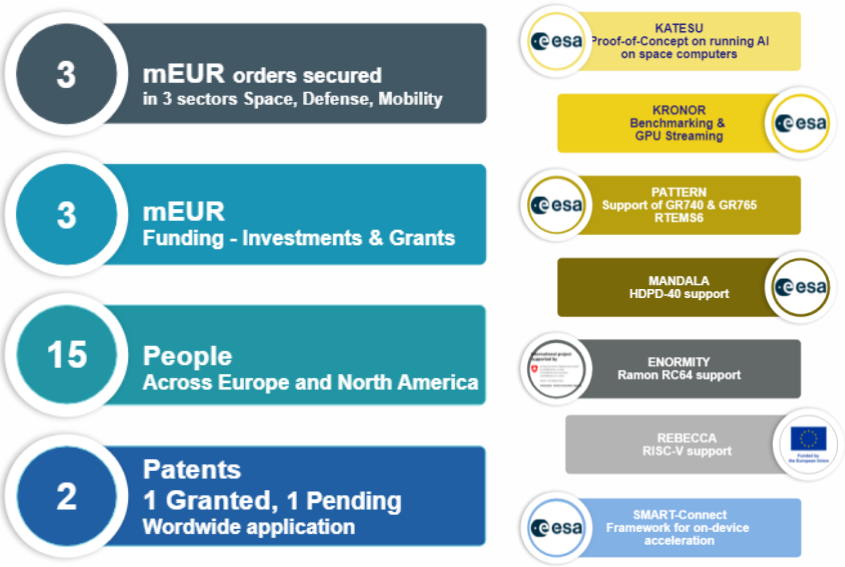
Our unified API for AI acceleration enables a more vendor-independent approach for acceleration and better performance compared to generic solutions such as TensorFlow Lite. Most acceleration frameworks are processor/vendor specific and thus, if it becomes necessary to change the processor, all code and adaptation

have to be re-developed for the new hardware platform. This results in significant lack of flexibility and lock-in with vendor systems. By using the Klepsydra framework, this processor lock-in is avoided.

About Klepsydra

Klepsydra Technologies AG is a Swiss Small Business, founded in 2018, focusing on providing advanced, patented edge AI software solutions for industries such as aerospace, robotics, Smart Sensors and industrial automation. We focus on developing software that maximizes the potential of AI and edge data processing to deliver high performance and efficiency.
Our mission is to offer scalable and efficient software that integrates seamlessly across a wide range of hardware, from cost-effective processors to high-performance accelerators. By addressing the balance between computational power and efficiency, we help our clients enhance their operations and achieve their objectives.
Klepsydra Technologies aims to be a global leader in AI and edge data processing solutions. Our partnerships with edge processor manufacturers and operating system providers strengthen this vision. Our software has also been successfully tested in space through multiple projects, demonstrating its reliability and innovation.

Klepsydra Technologies has carried out several successful projects with ESA, European Space and IoT companies in the field of flight software and onboard Artificial Intelligence software in Space qualified computers applied to real Space applications.



Klepsydra has extensive knowledge of high performance programming of edge device and implementation of AI inference engines and associated models on embedded systems.

Contact

Klepsydra Technologies AG

6300 Zug, Switzerland

www.klepsydra.com

sales@klepsydra.com