European Conference on EDGE AI Technologies and Applications - EEAI

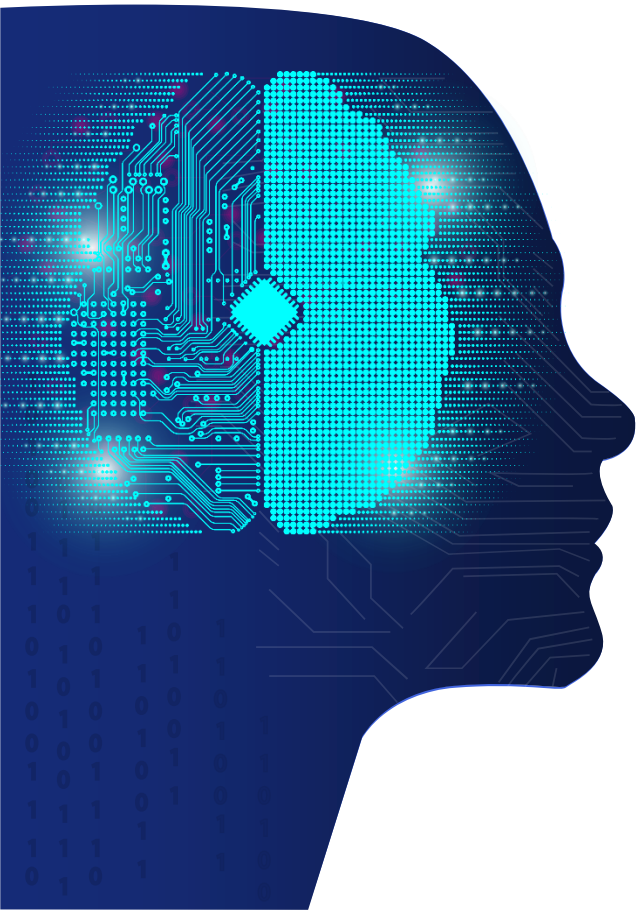
A Low Power And High-Performance Software Approach to Artificial Intelligence On-Board

Dr Pablo Ghiglino, Klepsydra Technologies, Switzerland

17-19 October 2023 Athens, Greece

Copyright © 2023

Presentation Outline



- Introduction
- State-of-the-art
- Research Goal
- Methodology
- Design and Implementation
- Experimental Results
- Future Research
- Conclusion

Copyright © 2023

Introduction

- Space onboard computers are unable to cope with the increasing volume of data and even less with the power consumption limits

Data generated at the edge grow much faster than the ability of edge computers to handle incoming data

Adding more computing power is either costly, requires too much energy or is limited by weight and volume

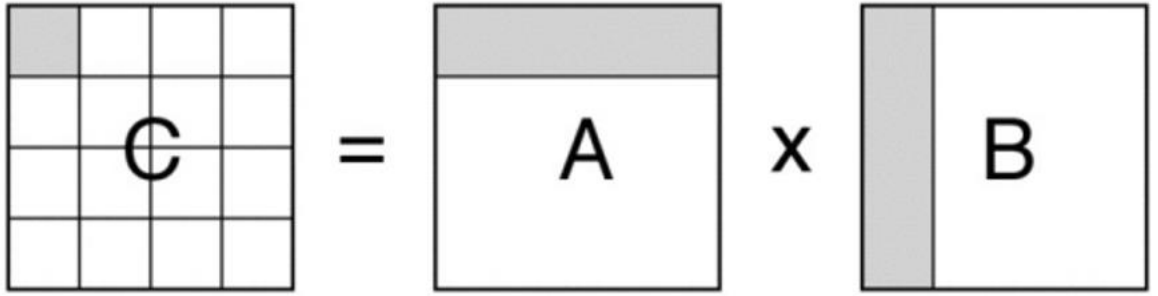
Huge efforts needed to optimise Bandwidth, Latency, Economics, Reliability / Resilience and Privacy / Security on the edge

Copyright © 2023

State-of-the-art

- The focus of state-of-the-art AI inference is parallelization of mathematical operations

Parallelisation



Parallel matrix multiplication: Each cell of the output matrix can be calculated independently of others, allowing high degree of parallelisation.

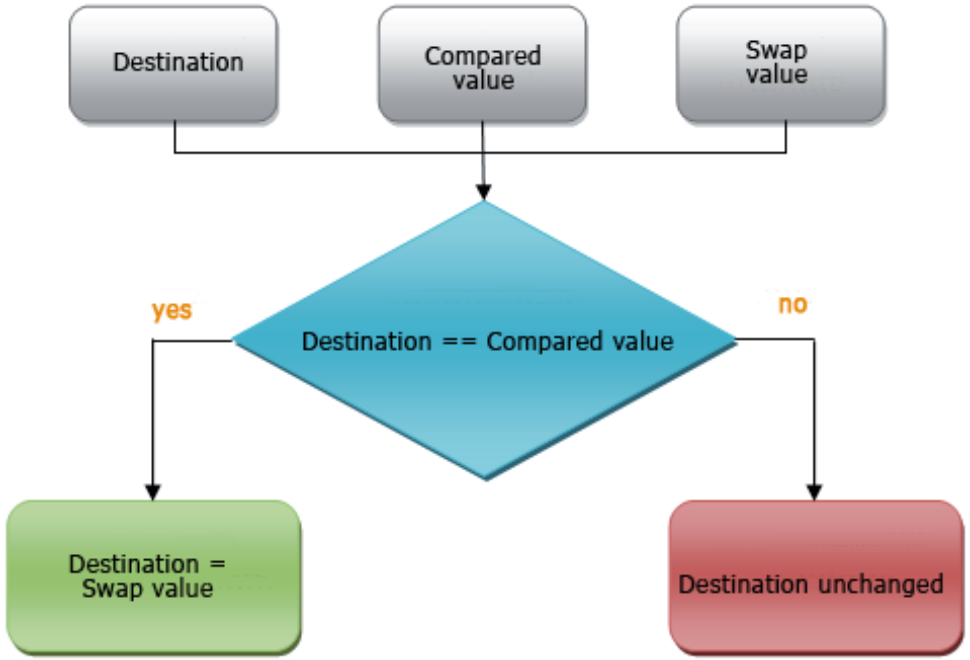
OpenMP

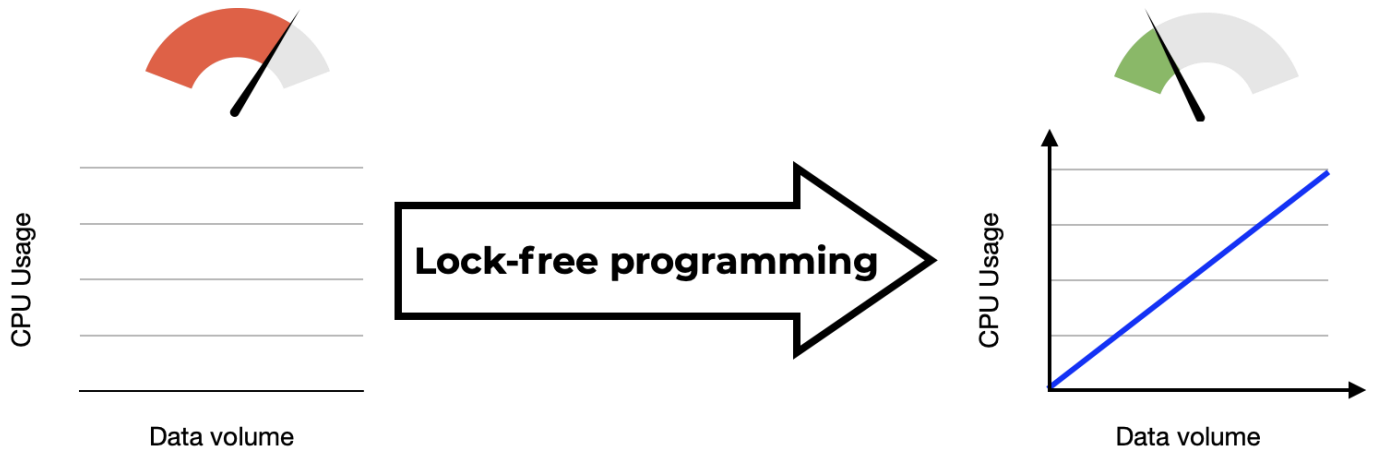
TensorFlow Lite

Copyright © 2023

Research Goal

- Application of lock-free programming to accelerate AI inference





- Compare-and-Swap has been an integral part of the IBM 370 architectures since 1970.
- Maurice Herlihy (1991) proved that CAS can implement more of these algorithms than atomic read, write, and fetch-and-add

Copyright © 2023

Methodology

- Lock-free programming can enable pipelining efficiently

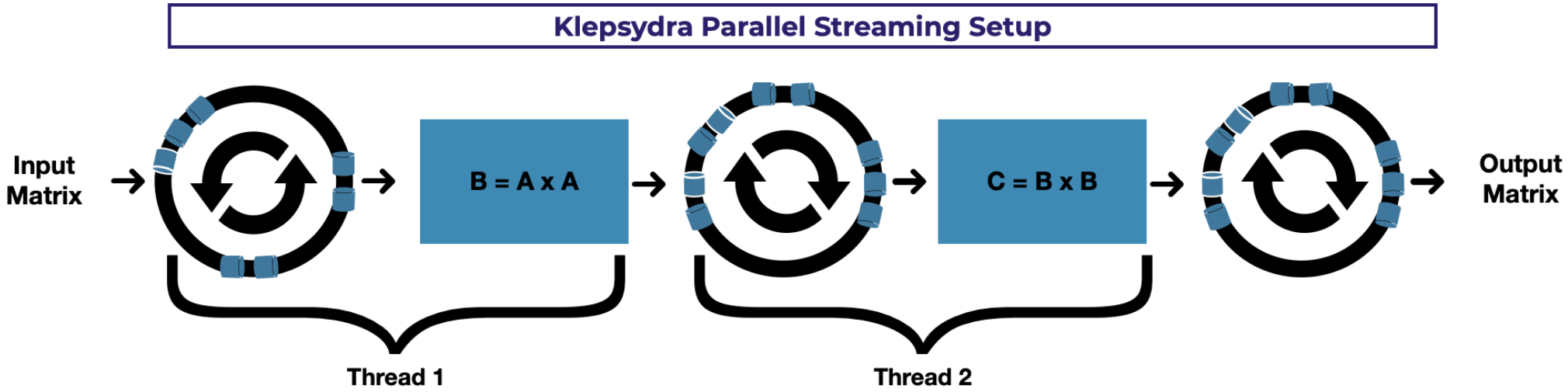
Parallelisation

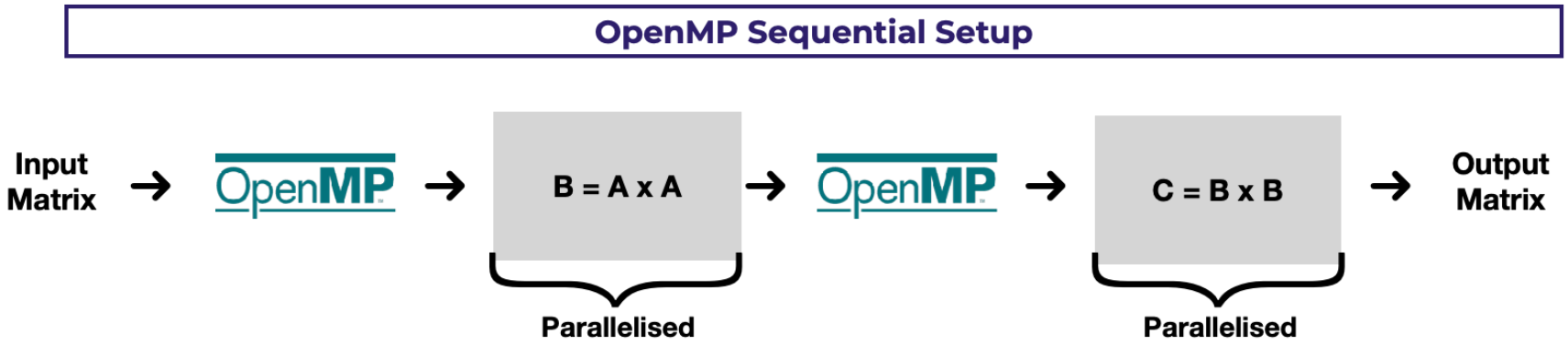
Pipeline

Copyright © 2023

Design and Implementation

- Usage our SDK ring-buffers for pipelining: **the eventloop**
- Preliminary testing of pipelining using matrix multiplication

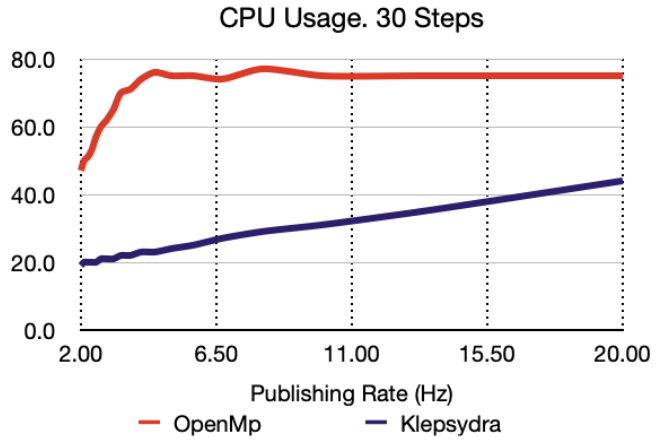


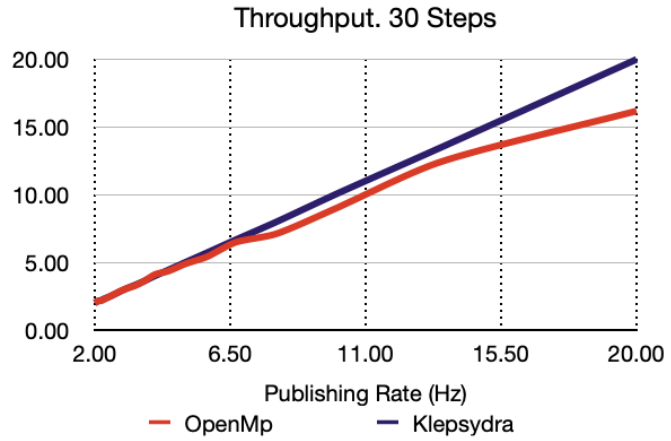


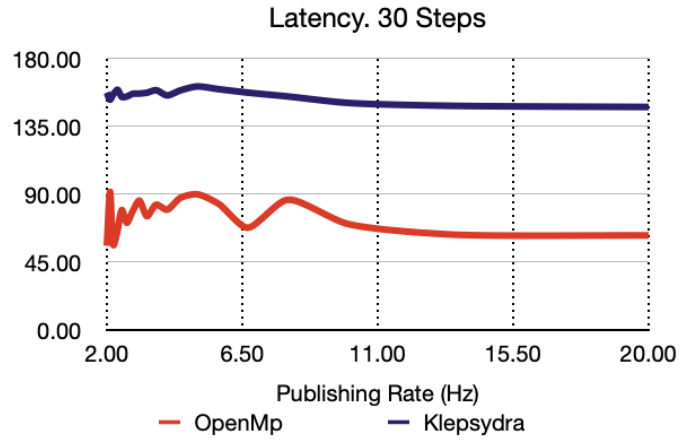
Copyright © 2023

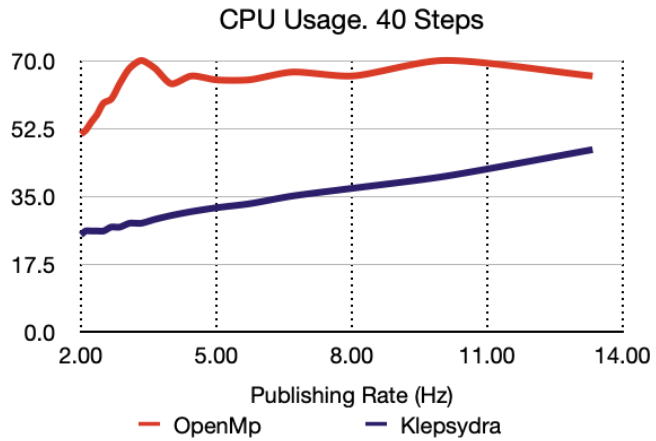
Design and Implementation

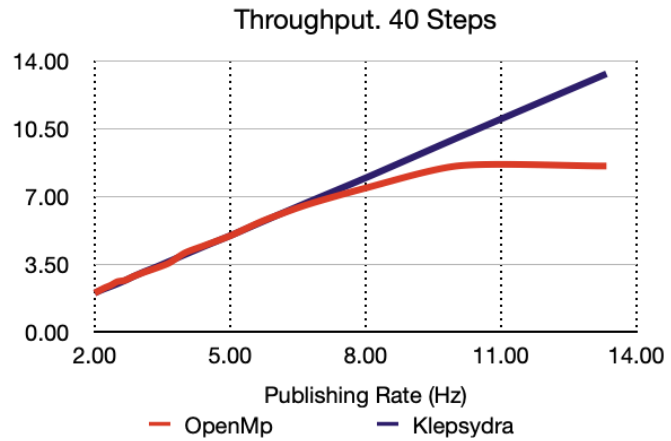
- Promising performance results: for increasing complexity, pipelining has a better throughput and lower CPU usage.

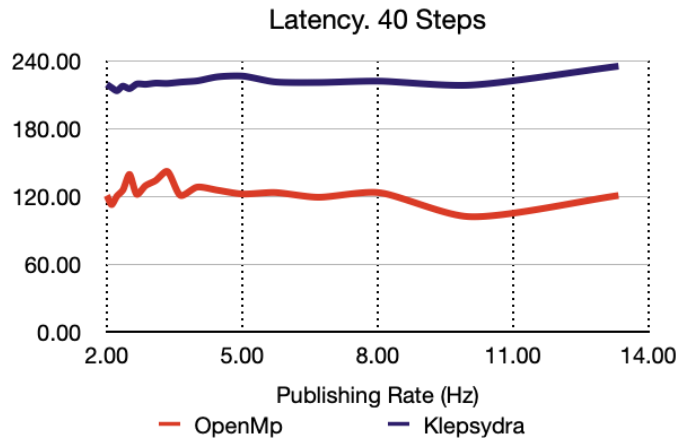








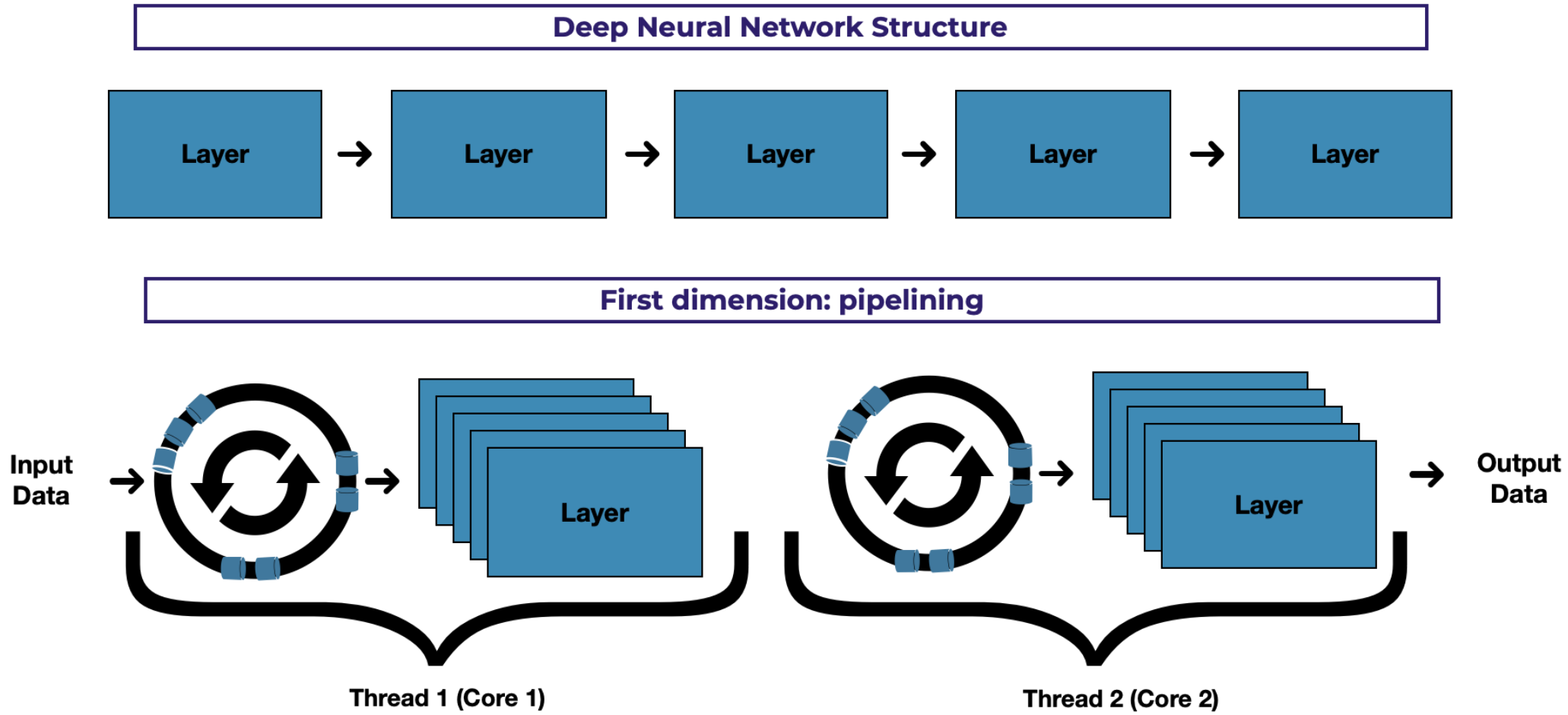




Copyright © 2023

Design and Implementation

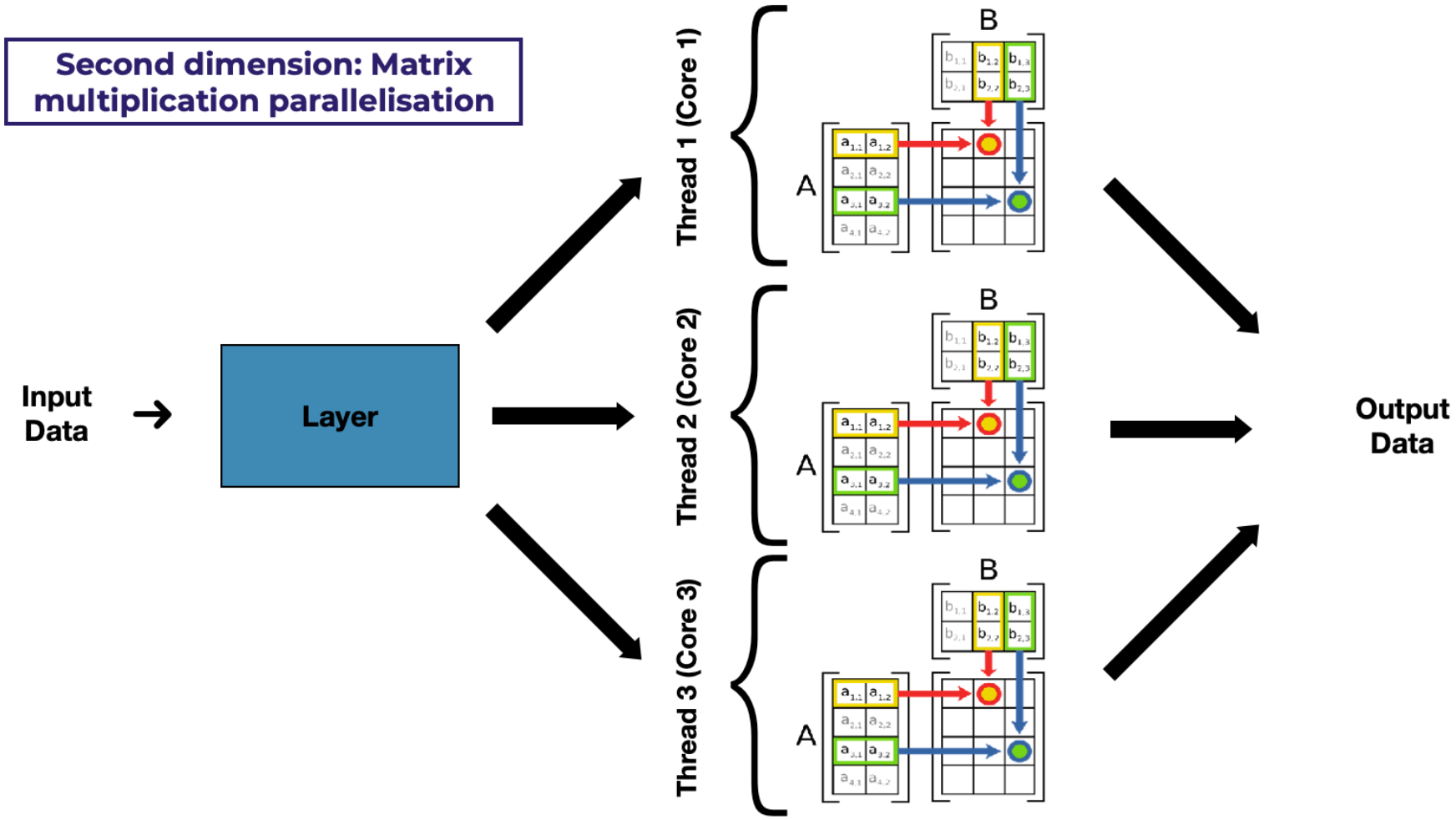
- The proposed solution consists in a 2-dim threading model: pipelining and parallelization: first dim below:



Copyright © 2023

Design and Implementation

- The second dimension is traditional parellisation

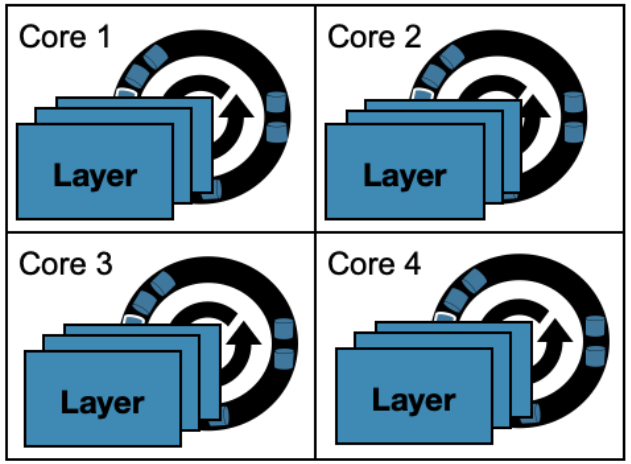


Copyright © 2023

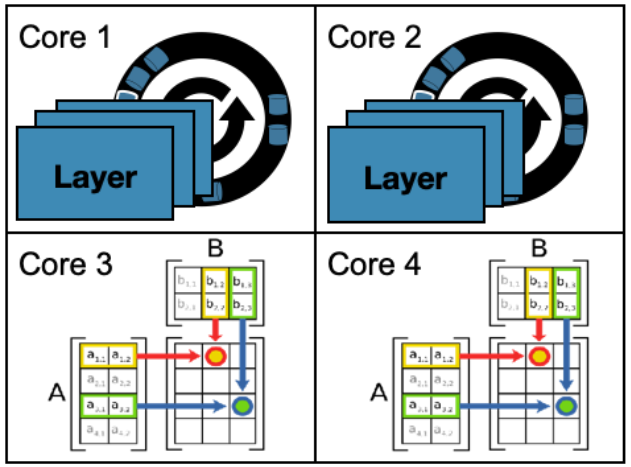
Design and Implementation

- Empirically we realized that different threading configurations lead to different performance criteria to be optimal:

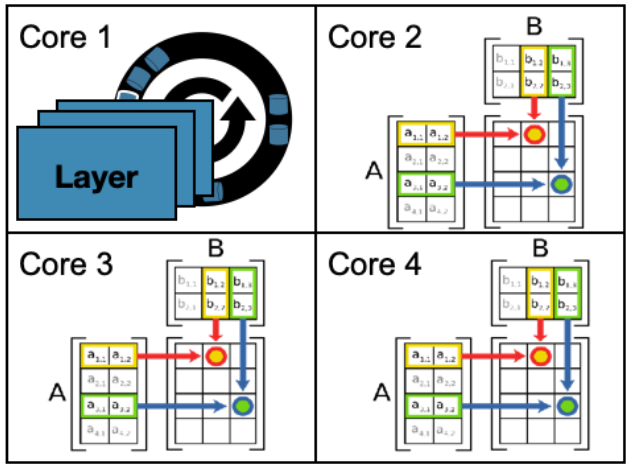
Threading model configuration modes



- **Low CPU**
- **Mid throughput CPU**
- **High latency**



- **Mid CPU**
- **High throughput CPU**
- **Mid latency**

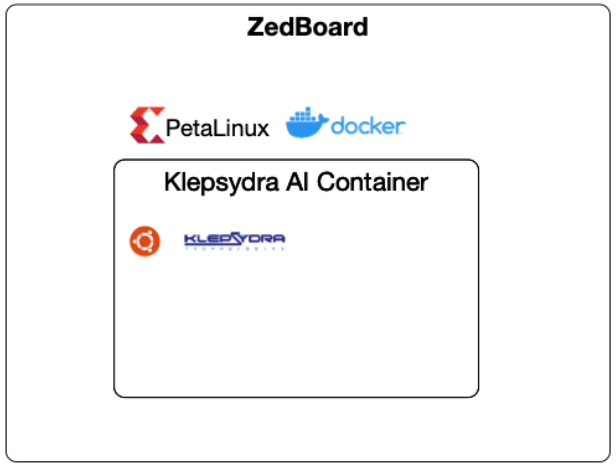


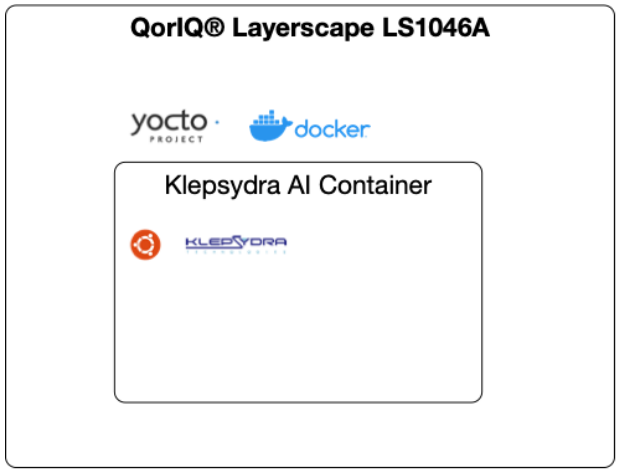
- **High CPU**
- **Mid throughput CPU**
- **Low latency**

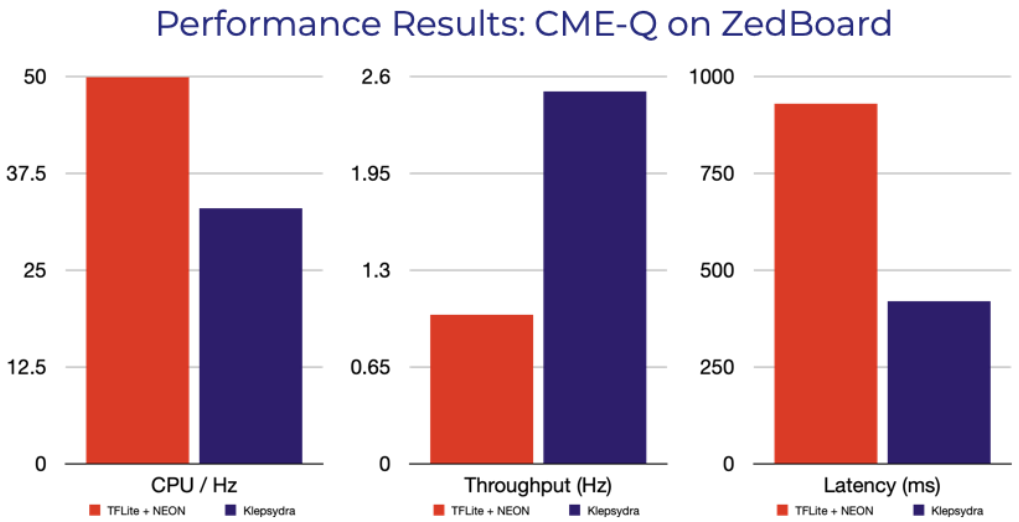
Copyright © 2023

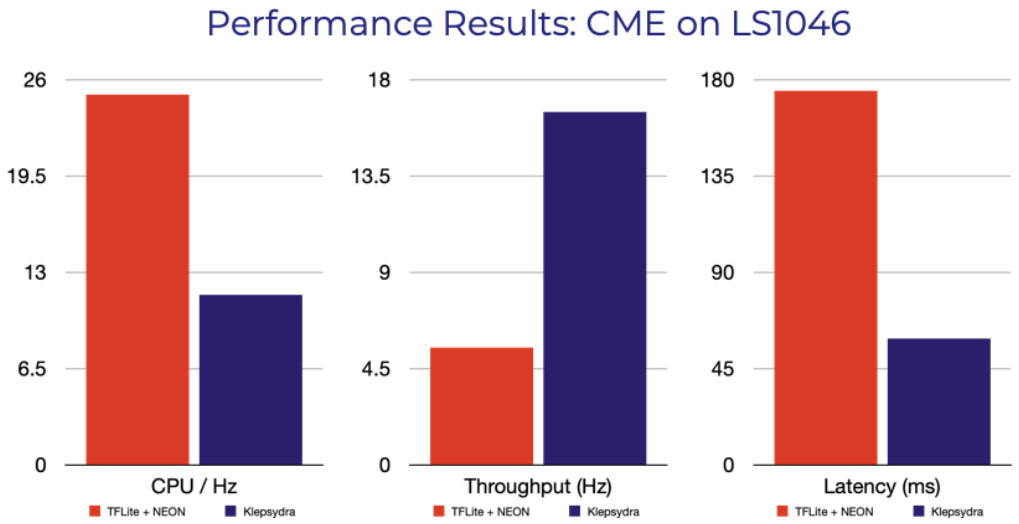
Experimental Results

- **The KATESU Contract:**
- Internal SWT Project
- Porting of Klepsydra AI to space applications:
  - Adoption of Klepsydra to Teledyne e2v LS1046 and ZedBoard
  - Using docker containers
  - Running Space application AI algorithms





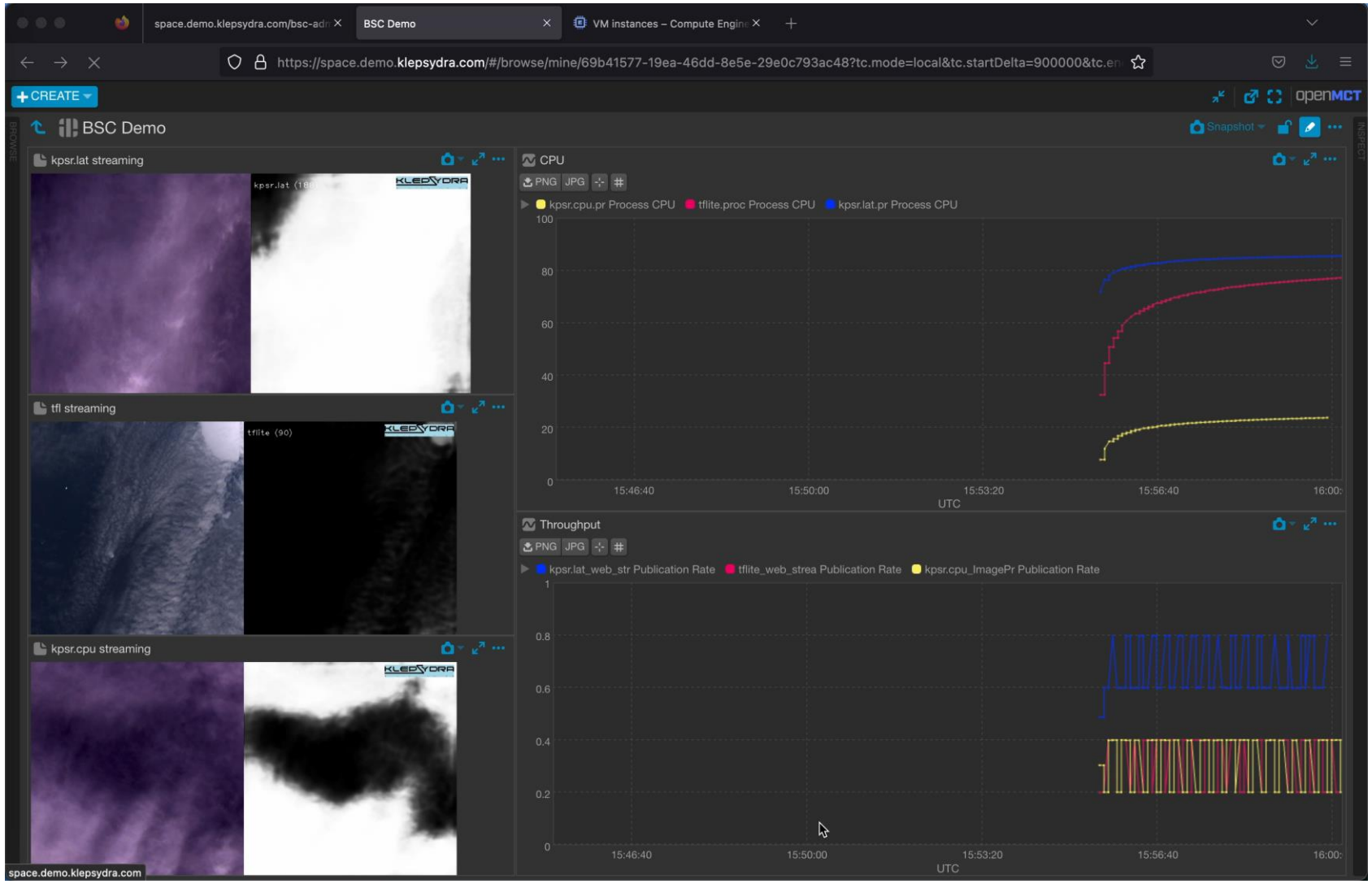




Copyright © 2023

Experimental Results

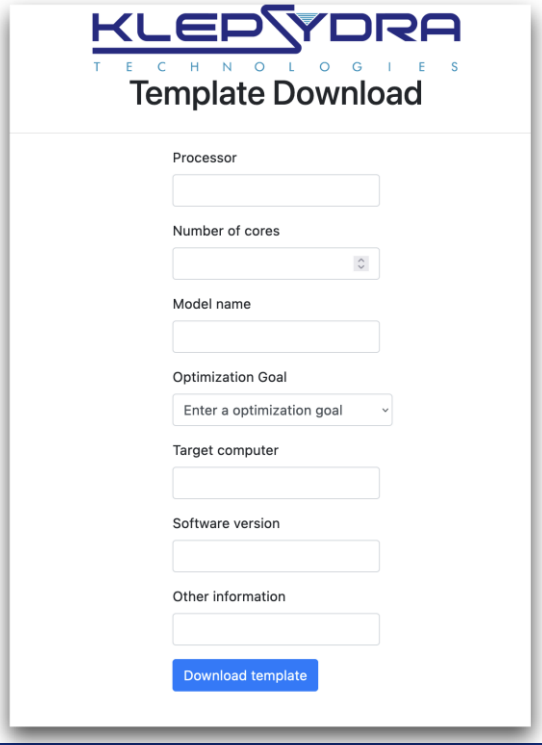
- **The Cloud Detection Demo:**
- Part of ESA's OBPMark-ML project
- Developed by BSC
- Demo shows the comparison of:
  - Klepsydra min latency
  - Klepsydra min CPU
  - Tensor Flow Lite

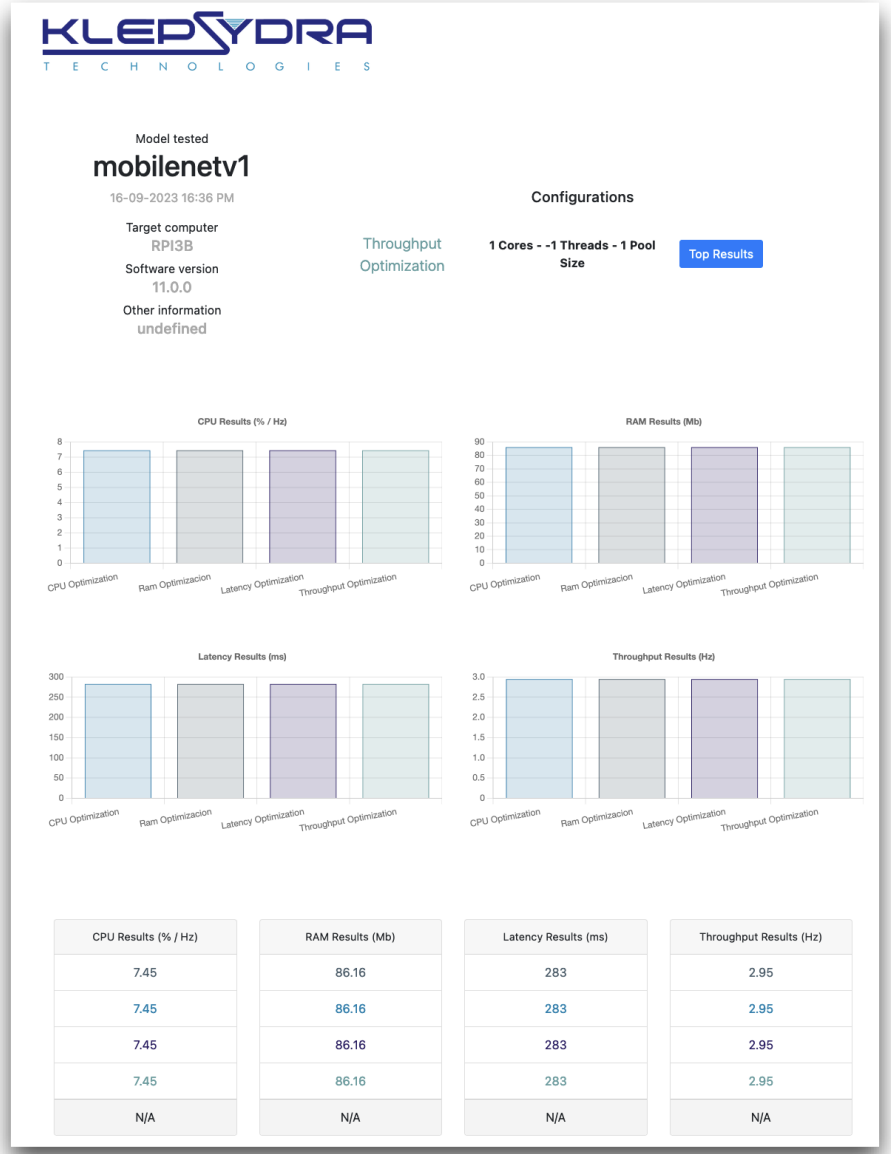


Copyright © 2023

Future Research

- Use of Convex Optimisation for optimal threading model search (right-side figures)
- Extension to hardware acceleration: FPGAs and GPU
- Extension to other architectures: RISC-V
- Porting to radiation hardened CPUs: GR765, GR740
- Extension to radiation hardened accelerators: HPDPv40 and RC64





Copyright © 2023

Conclusion

- The developed technology has substantial benefits for the space sector including low power consumption and high data rate:

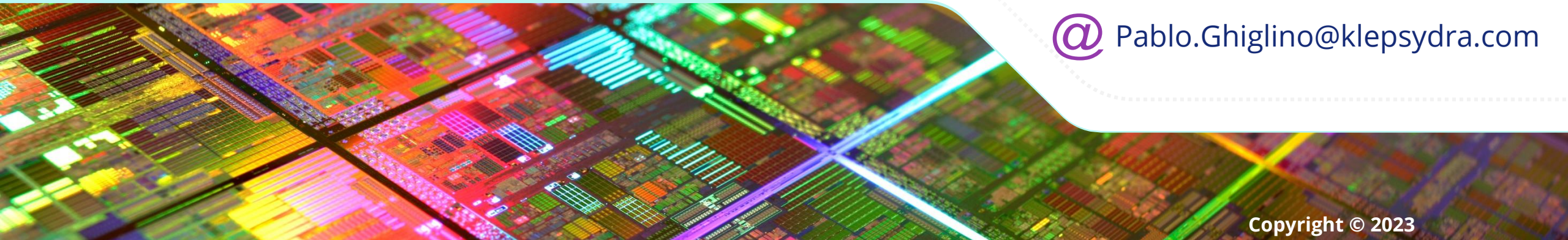
Vision-based navigation	Earth Observation	Telecommunications
• Process more images per second • Increase confidence in the mission	• Reduce power consumption up to 50% • Faster access to data from Earth	• Increase processed request per second (increase revenue) • Enable AI telecomm (Cognitive radios)

Copyright © 2023

EDGE AI

Thank You

For your attention



@ Pablo.Ghiglino@klepsydra.com

Copyright © 2023