HEIG VD | REDS Institut Reconfigurable and Embedded Digital Systems

RISC-V Summit

Hes·so

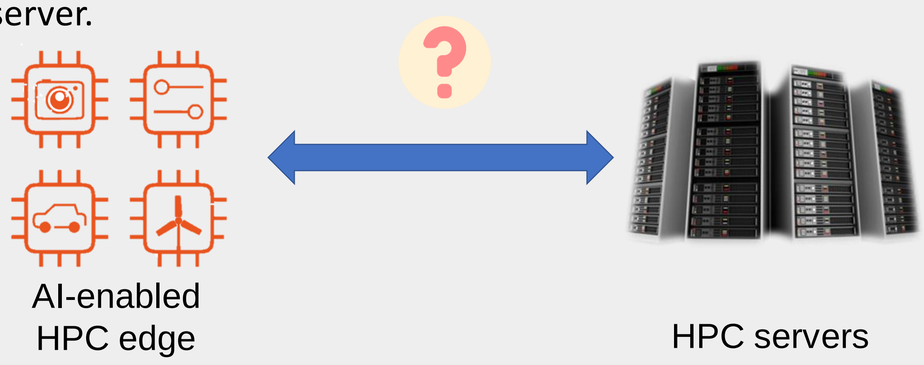
AN OPEN SOURCE FRAMEWORK FOR EDGE-TO-CLOUD INFERENCE ON RESOURCE CONSTRAINED RISC-V SYSTEMS

Mehdi Akeddar, Thomas Rieder, Guillaume Chacun, Bruno Da Rocha Carvalho and Marina Zapater

mehdi.akeddar@hes-so.ch

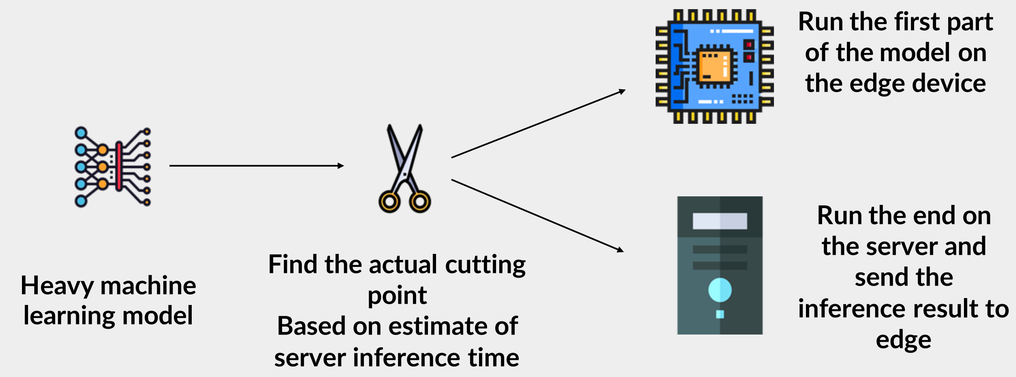
MOTIVATION AND CONTRIBUTIONS

- Current RISC-V based edge devices struggle to meet the performance requirements of complex AI inference.
- Efficient edge-to-cloud workload management is required.
- We propose a partial inference framework and algorithms to decide workload partitioning
- Real setup that automates the partitioning of ONNX and TFLite Models between a RISC-V accelerated nanodrone and a cloud server.

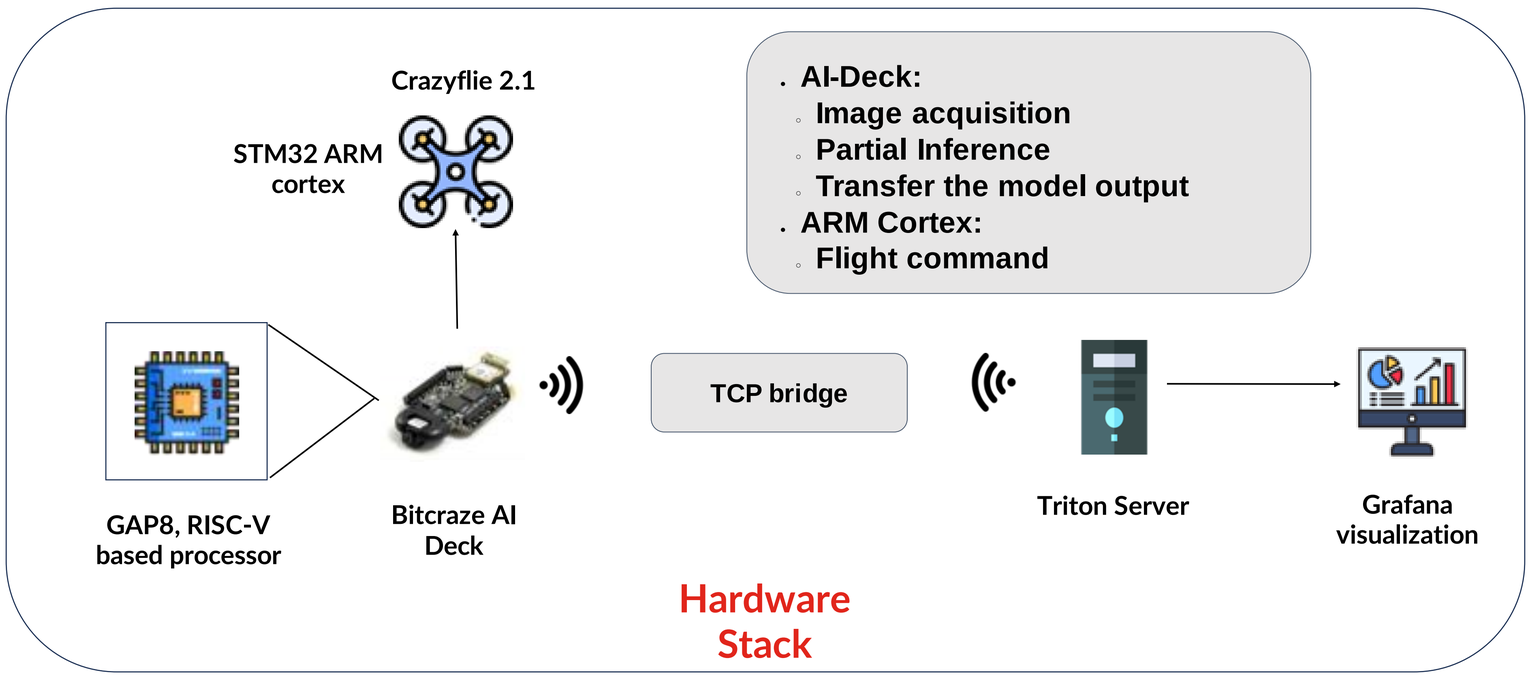


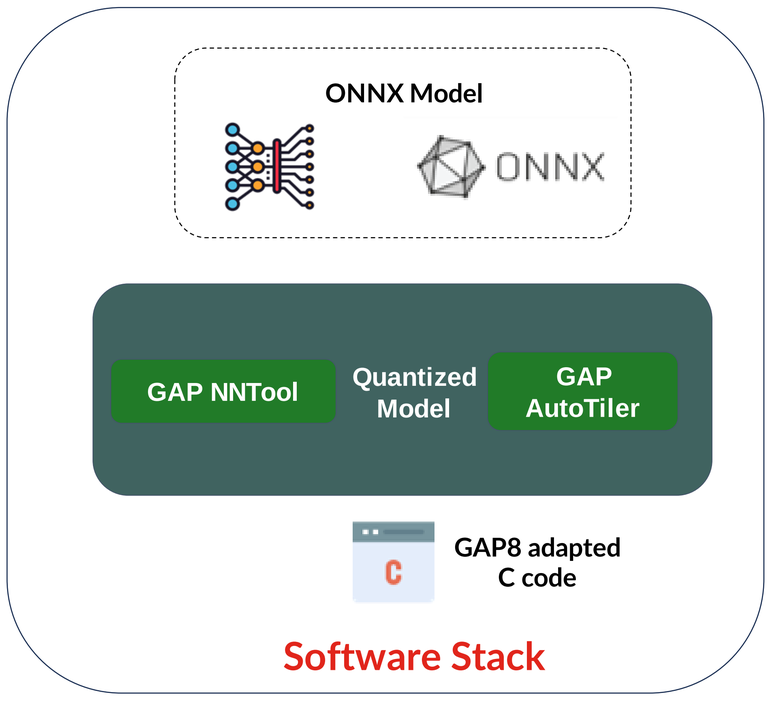
AUTODIDACTIC NEUROSURGEON

- ModifiedA modified version of the online autodidactic neurosurgeon algorithm [1] determines the optimal cutting points between edge and cloud inference.
- Considers both the model's architecture and external factors such as network conditions and server load.
- Predicts the time required for data transfer and server inference to select the cutting point.

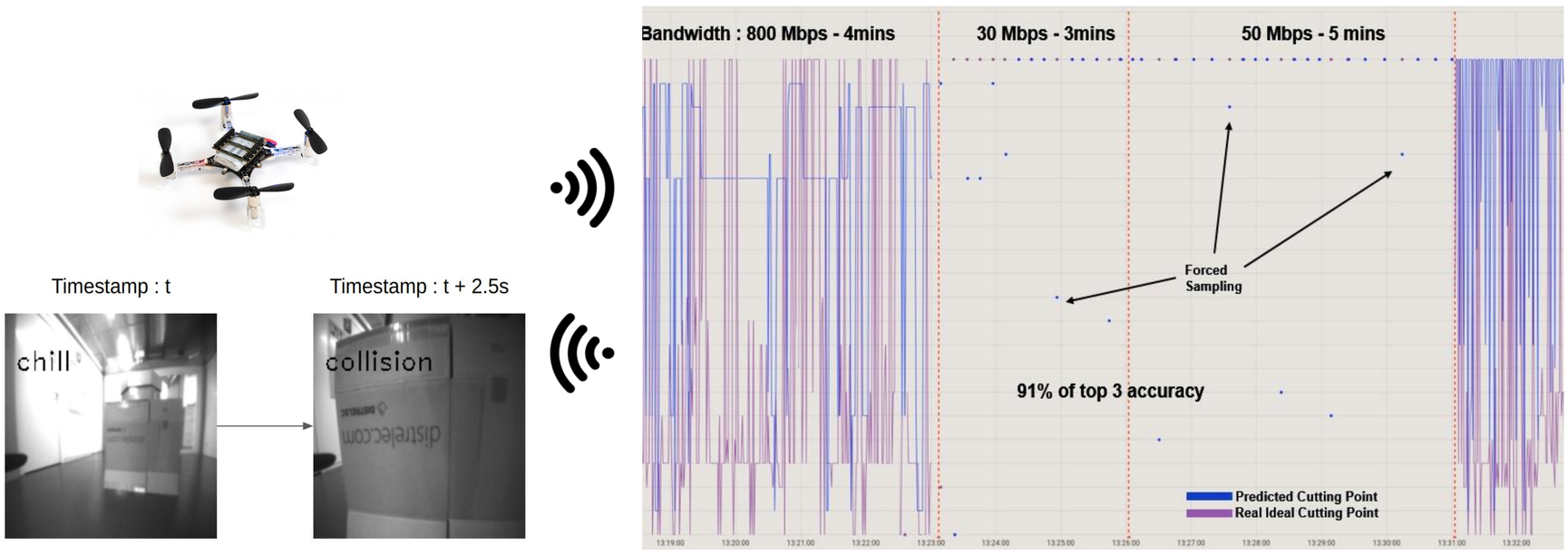


THE ECO4AI FRAMEWORK





RESULTS



CONCLUSIONS

- **Partial inference improved performance, encouraging the use of such setup**
- **Our framework will be released open-source as an out-of-the-box solution**
- **It is currently used for teaching a BSc course on autonomous navigation using the Crazyflie**

ACKNOWLEDGEMENTS

This work has been funded by the ECO4AI project granted by HES-SO, under the call for young researchers 2021.

REFERENCES

1. Letian Zhang, Lixing Chen, and Jie Xu. “Autodidactic neurosurgeon: Collaborative deep inference for mobile edge intelligence via online learning”. In: The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021 (Apr. 2021), pp. 3111–3123. doi: 10.1145/3442381.3450051. url: https://dl.acm.org/doi/10.1145/3442381. 3450051. 16
2. Palossi et al. “A 64mW DNN-based Visual Navigation Engine for Autonomous Nano-Drones”. In: IEEE IoT (2018).