European Conference on EDGE AI Technologies and Applications - EEAI

Accelerating Magnetic Resonance Imaging on Heterogeneous Edge Compute Platforms

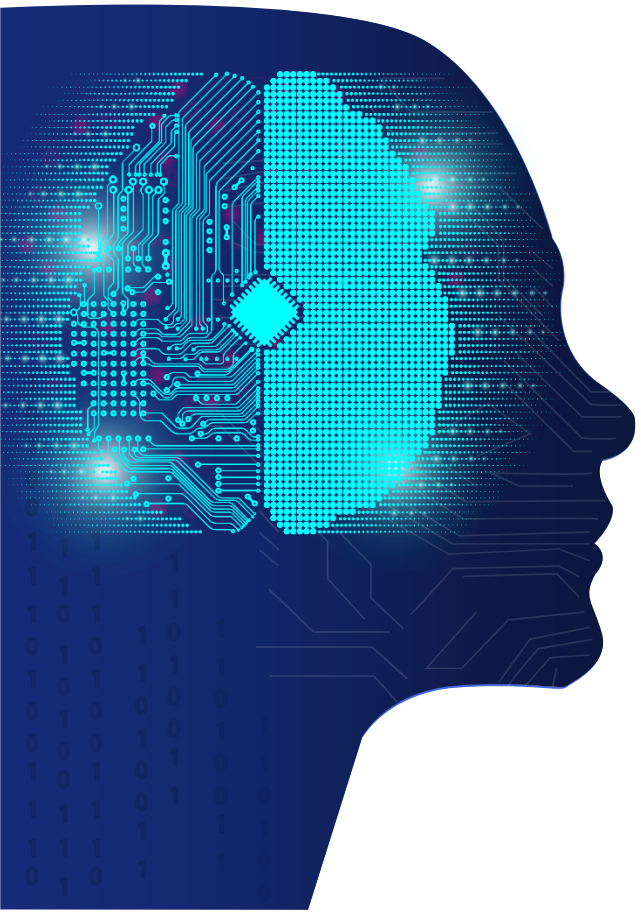
Yousef Alnaser, Fraunhofer ENAS, Ge
yousef.alnaser@enas.fraunhofer.de

Fraunhofer ENAS

17-19 October 2023 Athens, Greece

Copyright © 2023

Presentation Outline

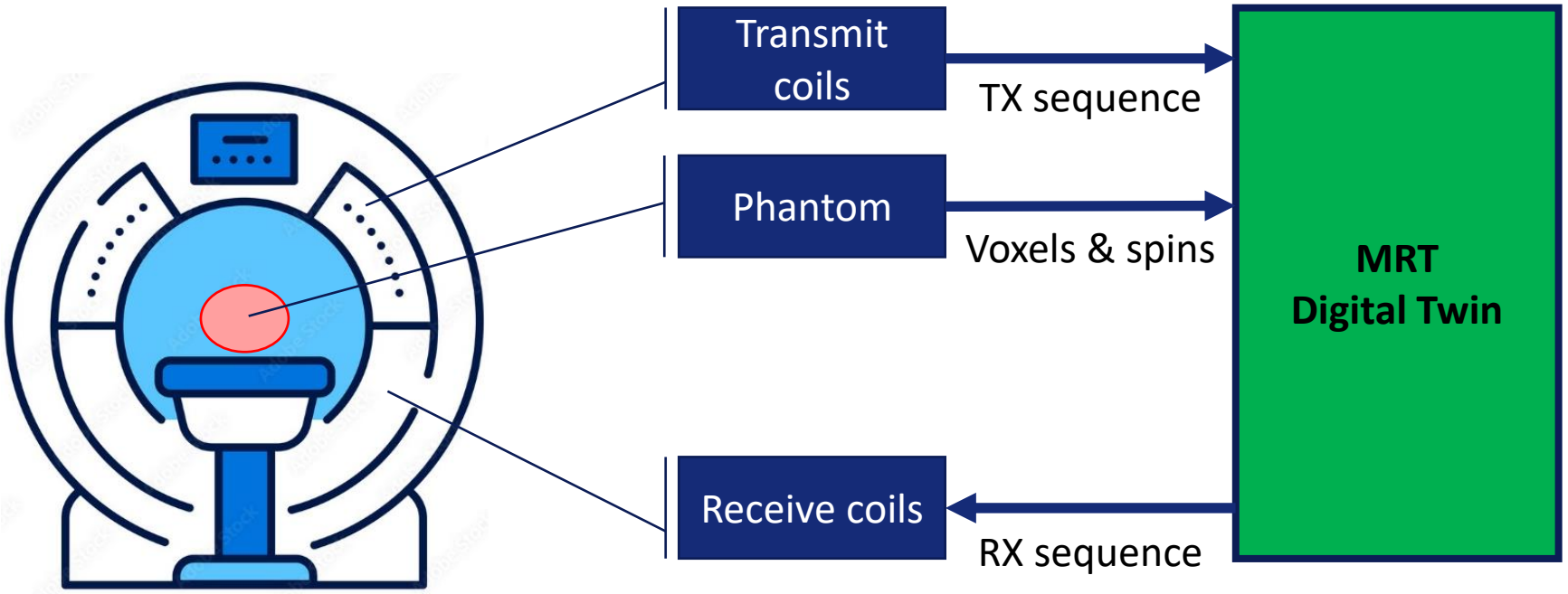


- Introduction
- State-of-the-art
- Research Goal
- Methodology
- Design and Implementation
- Experimental Results
- Future Research
- Conclusion

Copyright © 2023

Introduction

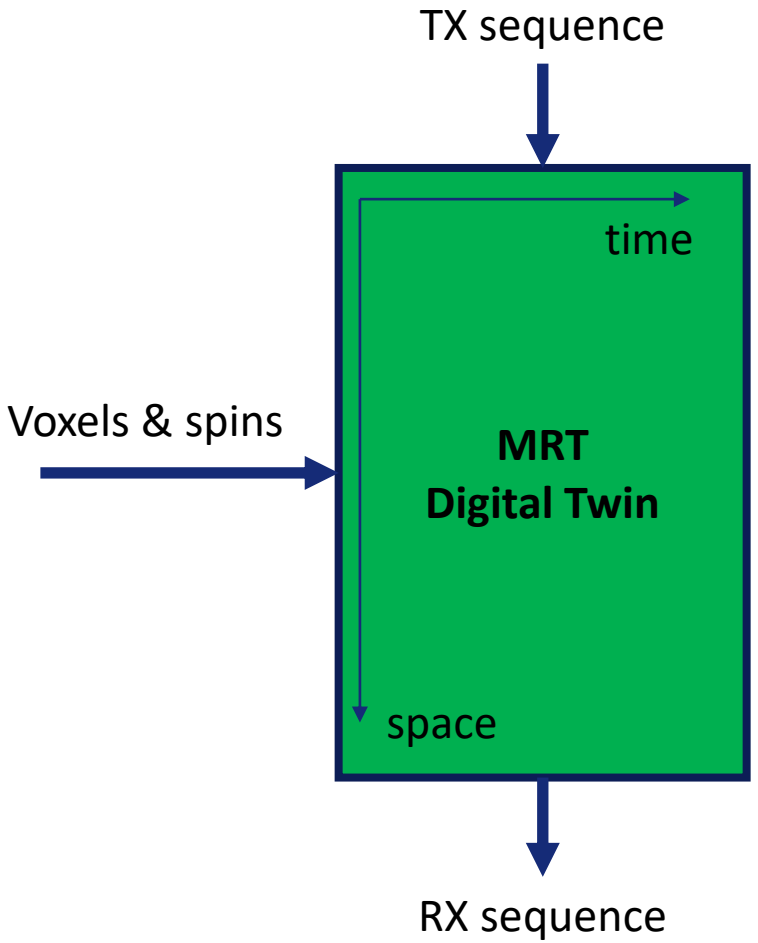
- Accelerate a real-time digital twin of the physical MRI process
- Added value delivered to customers
- Edge acceleration hardware needed (“thick edge”)



Copyright © 2023

State-of-the-art

- Algorithm complexity: $\mathcal{O}(T \cdot |spins|)$
- Compute requirement:
  - 1 MHz × 100 million spins = 100 Tspins/s
- Customer has a Matlab simulation based on solving the Bloch equation $\frac{d}{dt}\begin{pmatrix} M_x \\ M_y \\ M_z \end{pmatrix} = \begin{pmatrix} -\frac{1}{T_2} & \lambda B_z & -\lambda B_y \\ -\lambda B_z & -\frac{1}{T_2} & -\lambda B_x \\ \lambda B_y & -\lambda B_x & -\frac{1}{T_1} \end{pmatrix} \begin{pmatrix} M_x \\ M_y \\ M_z \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \frac{M_0}{T_1} \end{pmatrix}$



Acceleration by a factor 1.6 million needed

Copyright © 2023

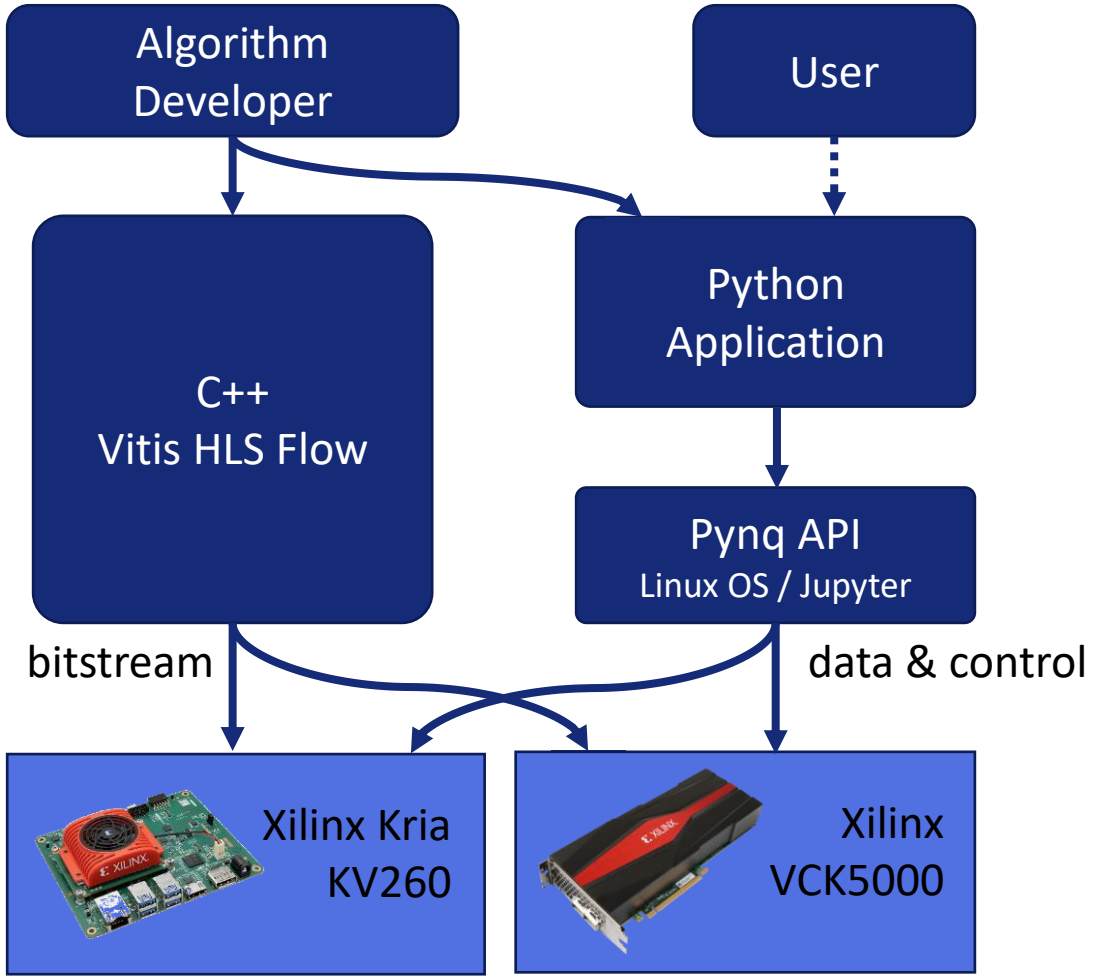
Research Goal

- Edge acceleration solution
  - Deployed with the equipment – no cloud solution!
  - Type of compute not very well suited for GPUs
- Huge compute capacity
  - Up to PFLOPS
  - Manageable energy consumption and price
- Flexibility
  - Algorithm will be adapted during life time
  - Features shall be switched on and off
- Easy to program
  - By a software programmer, not a hardware developer

Copyright © 2023

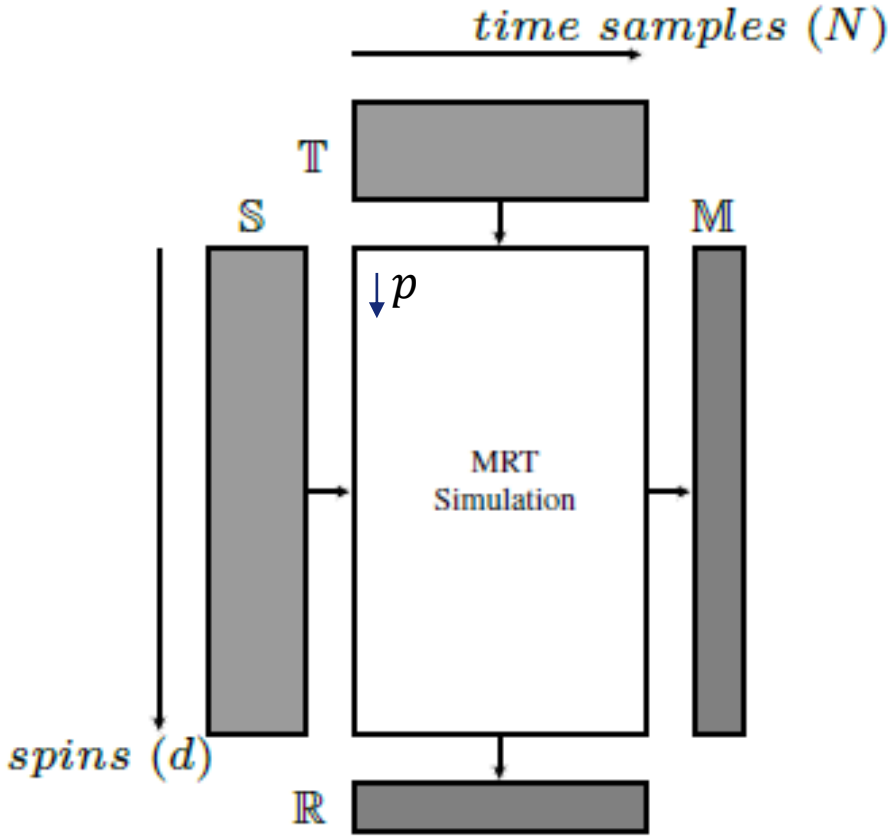
Methodology

- Well-established design methodology (10+ years)
  - Still not wide-spread in industry
- No explicit hardware development with High-level synthesis (HLS)
  - Drawbacks exist ☺
- Fast design space exploration and turnaround time
- Easy integration in application code
  - Only accelerate the slow parts
- Scalable to different acceleration boards



Copyright © 2023

Design and Implementation

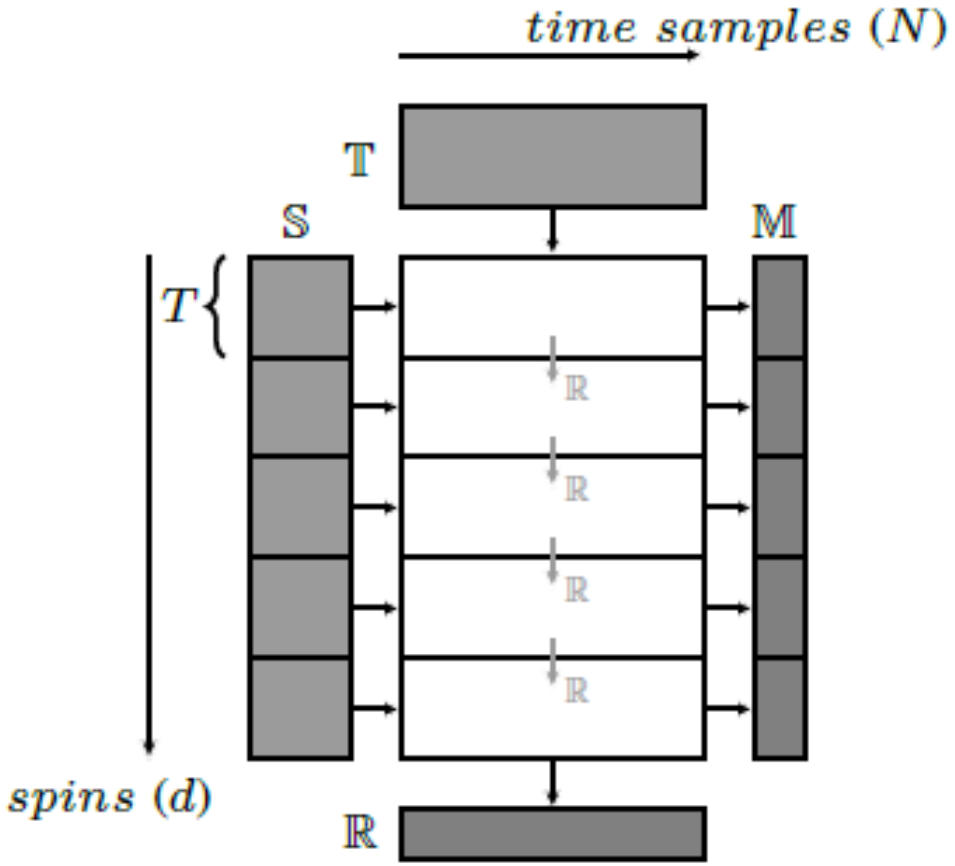


- Solving for each time step and each spin:
  - $$\begin{pmatrix} M_{xy_out} \\ M^*_{xy_out} \\ M_{z_out} \end{pmatrix} = \begin{pmatrix} (\alpha^*)^2 & -\beta^2 & 2\alpha^*\beta \\ -(\beta^*)^2 & -\alpha^2 & 2\alpha\beta^* \\ -\alpha^*\beta^* & -\alpha\beta & \alpha\alpha^* - \beta\beta^* \end{pmatrix} \begin{pmatrix} M_{xy_in} \\ M^*_{xy_in} \\ M_{z_in} \end{pmatrix}$$
- Data dependency in time (magnetization state)
- No data dependency in space
  - Result aggregates in space

➔ Parallelization in space – vector size p

Copyright © 2023

Design and Implementation

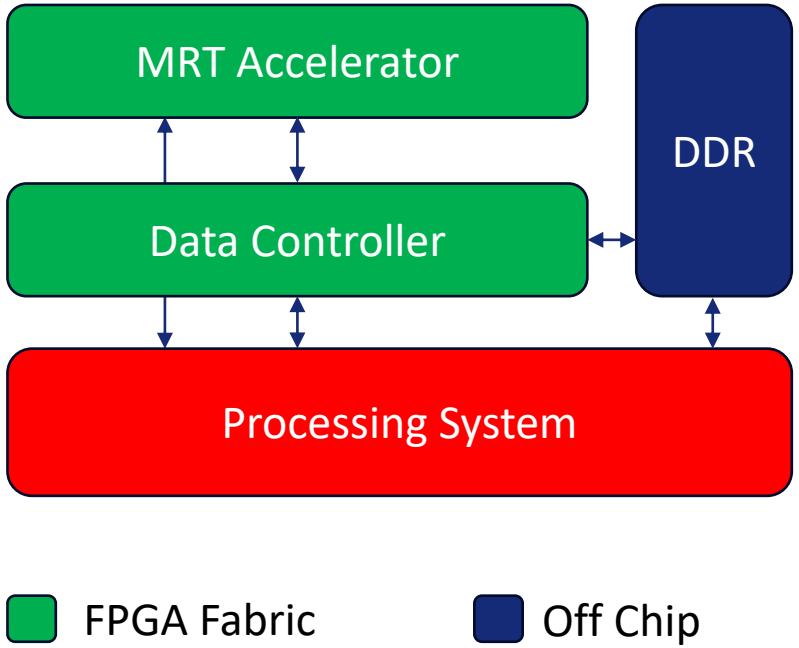


- Bandwidth problem:
  - $\mathbb{T}$ and $\mathbb{S}$ do not fit in on-chip memory
  - Naive implementation reloads $\mathbb{T}$ for $\frac{d}{p}$ times
- Solution:
  - Tiling of $\mathbb{S}$ in space
  - Select T such that the tile fits in local memory
  - Now $\mathbb{T}$ is reloaded only $\frac{d}{T}$ times

Copyright © 2023

Design and Implementation

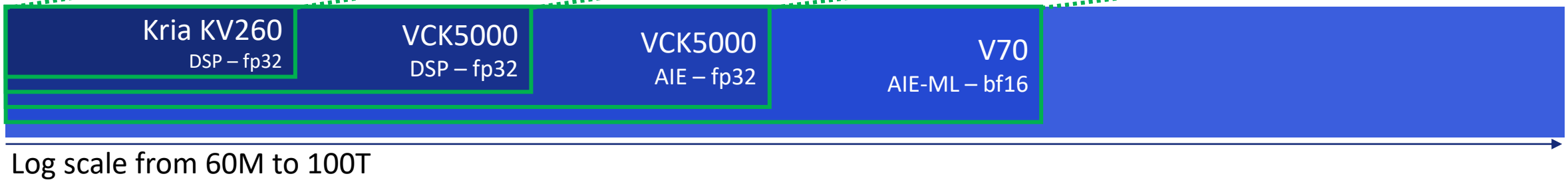
- Accelerator
  - 300 LoC of HLS C++ code
  - Controlled by the Pynq API on the ARM core
  - Pre and post-processing runs in Python
- Data Movement
  - Xilinx standard data mover IP
  - HLS component for buffer description generation
- Configurability
  - Parallelization factor p
  - Tiling size T



Copyright © 2023

Experimental Results

	Matlab	Kria KV260	VCK5000	VCK5000 (projected)	Alveo V70 (projected)
Performance in Spins/s	60 M	850 M	7.5 G	67 G	815 G
Resource Utilization	CPU	92% (DSP)	90% (DSP)	60% (AIE)	60% (AIE-ML, bfloat16)



- Existing results already very helpful to customer
- 125x faster simulation time

Copyright © 2023

Future Research

- Reduced precision
  - Single-precision at the moment
  - Going to half-precision promises 4x performance (e.g. bfloat16 on Alveo V70)
- Multiple boards
  - Scale out to 2, 4 or 8 boards
- Algorithmic improvements, e.g.
  - *K. Majewski, Rotation relaxation splitting for optimizing parallel RF excitation pulses with T1- and T2-relaxations in MRI, J. Magn. Reson., 288 (2018), pp. 43-57*
- Additional features
  - More flexible MRI equipment
  - More physical aspects

Copyright © 2023

Conclusion

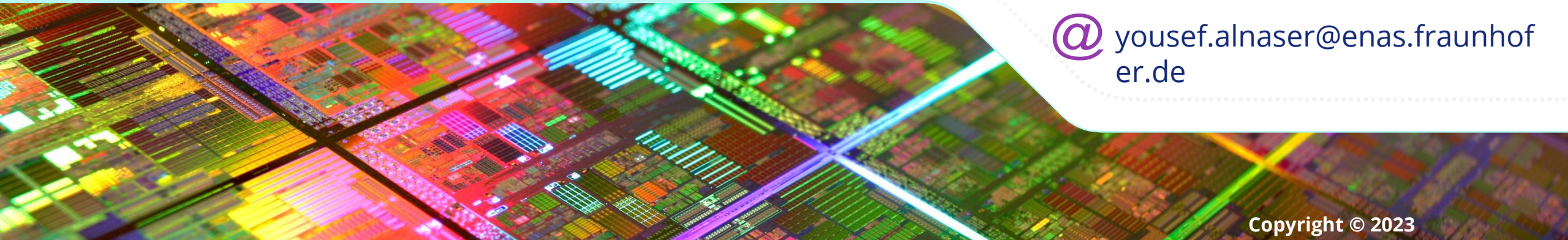
- Investigated accelerating MRI simulations solving the Bloch equations
  - Customer wants a real-time edge solution
  - Current Matlab solution is ~125,000 times too slow
- FPGA-based demonstrator on Kria and VCK5000
  - Achieved 125x speedup (VCK-DSPs)
  - Highly scalable solution
  - Software-programmable with C++ and Python
- Still factor 13,000 off the target speed:
  - AI acceleration cores (AIE) – 9x
  - Lower precision, bfloat16 on AIE-ML – 12x
  - Algorithmic improvements – 50x
  - Multiple boards – 4x

Copyright © 2023

EDGE AI

Thank You

For your attention



@ yousef.alnaser@enas.fraunhofer.de

Copyright © 2023