ZuSE

TECHNISCHE UNIVERSITÄT DRESDEN

On the ZuKiMo Chip AI Accelerator for Automotive

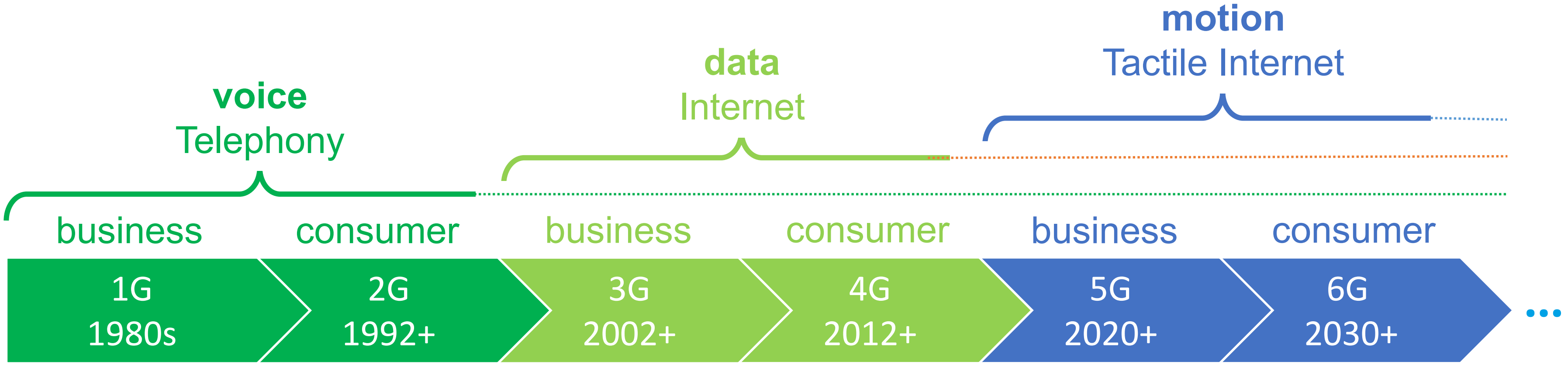
Gerhard Fettweis (presenting) – Vodafone Chair Professor @ TU Dresden; Scientific Director & CEO @ Barkhausen Institut

Major contributors: Simon Friedrich, Robert Wittig, Emil Matus, and more

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

6G Cellular Communications Poses a Need to Act



voice
Telephony

data
Internet

motion
Tactile Internet

business	consumer	business	consumer	business	consumer
1G 1980s	2G 1992+	3G 2002+	4G 2012+	5G 2020+	6G 2030+

…

entertainment
gardening
chores
mobility
tools
health
fitness, …

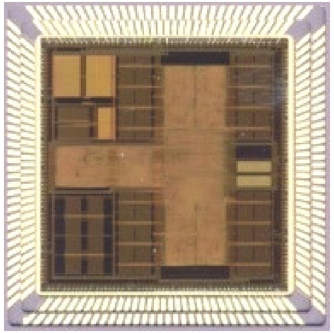
consumer
cordless

consumer
WiFi/laptop

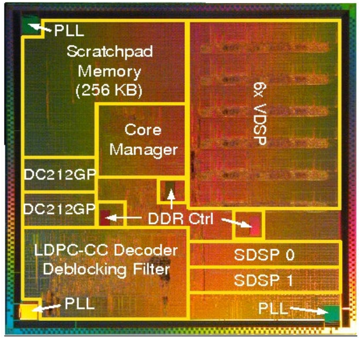
consumer
home robots

Platform Chips – Some of Our MPSoC Examples

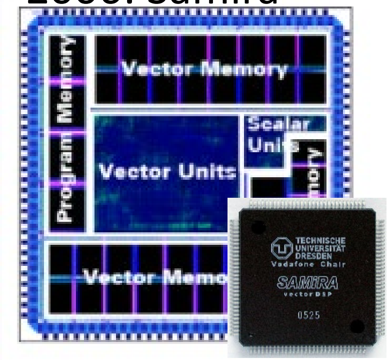
1997: M3 DSP



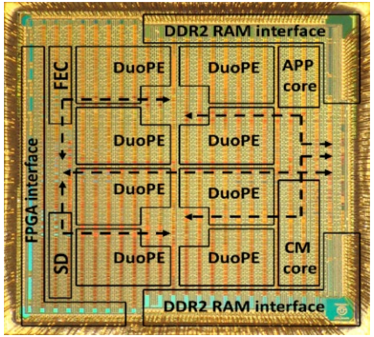
2004: Tomahawk 1



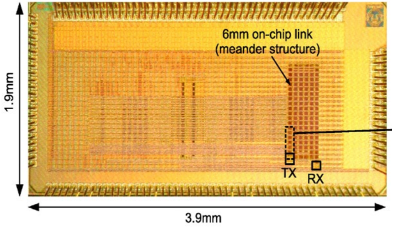
2006: Samira



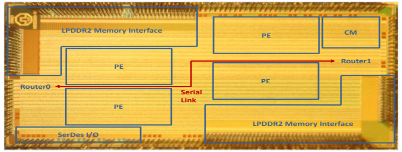
2008: Tomahawk 2



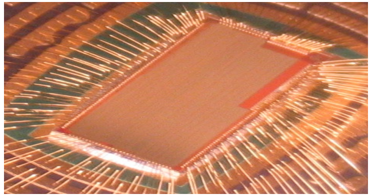
2012: Atlas



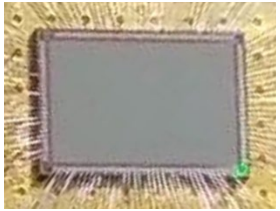
2014: Tomahawk 3



2016: Tomahawk 4



2019: Kachel



1997 M3-DSP	M. Hosemann, et al., "Applications for the Highly Parallel Mobile Multimedia Modem M3 DSP," in Proceedings of EUROMICRO Conference (Euromicro 2002), Dortmund, Germany
2004 Tomahawk1	P. Robelly, et al., "Implementation of Recursive Digital Filters into Vector SIMD DSP Architectures.," Proceedings of IEEE ICASSP 2004, Montreal, Canada, May 2004
2006 SAMIRA	E. Matus, et al. "A GFLOPS Vector-DSP for Broadband Wireless Applications," in Proceedings of IEEE Custom Integrated Circuits Conference (CICC 2006), San Jose, USA, Sep 2006
2008 Tomahawk2	T. Limberg, et al., "A Fully Programmable 40 GOPS SDR Single Chip Baseband for LTE/WiMAX Terminals," Proceedings of ESSCIRC 2008, Edinburgh, UK, Sep 2008
2012 Atlas	M. Winter, et al., "A 335Mb/s 3.9mm² 65nm CMOS Flexible MIMO Detection-Decoding Engine Achieving 4G Wireless Data Rates," (ISSCC 2012, San Francisco, USA, Feb 2012
2014 Tomahawk3	B. Nöthen, et al., "A 105GOPS 36mm2 Heterogeneous SDR MPSoC with Energy-Aware Dynamic Scheduling and Iterative Detection-Decoding for 4G in 65nm CMOS," ISSCC 2014, Paper 10.7
2016 Tomahawk4	S. Haas, et al., "An MPSoC for Energy-Efficient Database Query Processing," in Proceedings of Design Automation Conference (DAC 2016), Austin/Texas, USA, Jun 2016
2019 Kachel	G. Fettweis, et al., "A Low-Power Scalable Signal Processing Chip Platform for 5G and Beyond - Kachel," Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, California, USA

Challenge of a Future Digitization-Shaped Society:

Can we trust machines the same way?

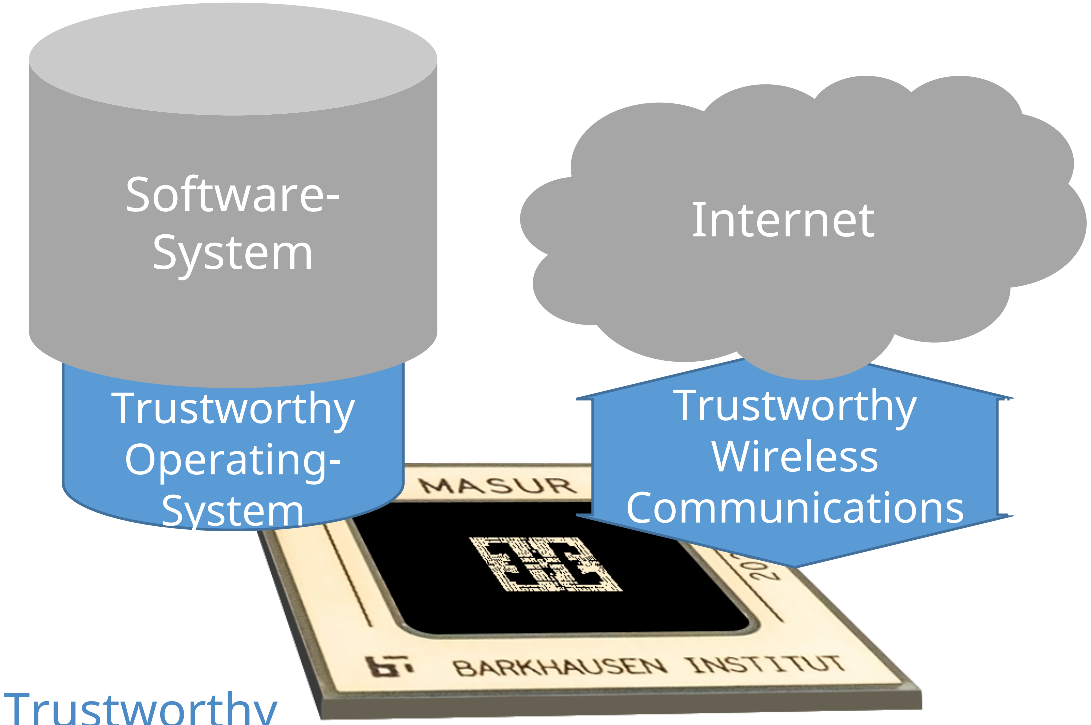
How do we trust humans?

What or whom exactly do we need to trust here?

How can we judge the trustworthiness?

Image copyright Shutterstock/Aleksandr Ozerov,
Shutterstock/Andrey_Popov

The BI Runs a Research Program on **Communications** and **Computing** Technology to Ultimatly Develop the **Methodology** Basis for a Trustworthy System Design.


Software-System
Trustworthy Operating-System
Internet
Trustworthy Wireless Communications
MASUR
BARKHAUSEN INSTITUT
202

Trustworthy Platform Chips

COREnext:
Europe’s Semiconductor Platform for 6G

Lead: Barkhausen Institute
Team: Ericsson, Nokia, Infineon, NXP, Sequans, Kalray, IHP, IMEC, LETI, Australo, TUD,…

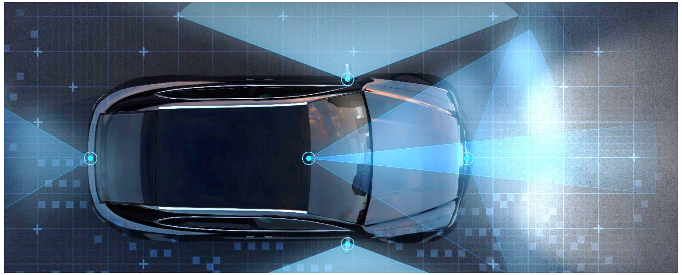
LinkedIn	/corenext-eu
Twitter	@COREnext_EU
Email	info@corenext.eu
Website	www.corenext.eu

Funded by
the European Union

TECHNISCHE UNIVERSITÄT DRESDEN

Autonomous Driving – Challenges

Automated Vehicles – What a Vision



What if we could get into a driverless car

- During a snow storm
- At night
- After dinner & drinking

How safe is safe

- AV cars cannot excuse themselves for errors
- AV cars must perform better than humans ➔ 10x
- 1/10 of lethal accidents compared with humans
- We can only decide after 1000 deaths by AVs if really safe…???

Automated Driving – The Huge Data Challenge:

How many times must one visit every meter of road to achieve a statistically significant reliable sample?
➔ only after approximately 100 deaths/country: drive style and road style differs from country-to-country
➔ assuming 10% fatality over human driving we must calculate road visits via 1000 human driven deaths

$$road\ visits = \frac{km\ driven}{death} \times \frac{1}{road\ length} \times 1000 = \frac{1000}{\frac{deaths}{km} \times road\ length}$$

	2023 Deaths / km	2023 Road length	Required Road visits
Germany	$1.65 \cdot 10^{-9}\ /km$	$0.6 \cdot 10^{6} km$	1m
Japan	$1.68 \cdot 10^{-9}\ /\mathrm{km}$	$1.28 \cdot 10^{6} km$	0.5m
USA	$8.6 \cdot 10^{-9}\ /km$	$6.59 \cdot 10^{6} km$	18k
China	$6.3 \cdot 10^{-5}\ /km$	$6 \cdot 10^{6} km$	2.6k

AI/ML – The Signal Processing Challenge

Signal Processing – We Love Linear (Matrix) Algebra

The minimum requirement for “linear Algebra” is a semiring:

Operator ① a semigroup over set $\mathfrak{D}$ e.g. ① is × multiplication, $\mathfrak{D}=\mathbb{N}$

Operator ② a semigroup over set $\mathfrak{D}$ e.g. ② is + addition, $\mathfrak{D}=\mathbb{N}$

Operator ① is distributive over ②

$$a \textcircled{1} \left(b \textcircled{2} c\right) = \left(a \textcircled{1} b\right) \textcircled{2} \left(a \textcircled{1} c\right) = a \textcircled{1} b \ \textcircled{2} \ a \textcircled{1} c$$

e.g. $a \times (b + c) = (a \times b) + (a \times c) = ab + ac$

Matrices: $Y_{k+1} = X_k + A_k Y_k \;\rightarrow\; Y_{k+2} = \underbrace{\left(X_k + A_k X_k\right)}_{\text{look-ahead}} + \underbrace{\left(A_k A_{k+1}\right)}_{\text{look-ahead}} Y_k$

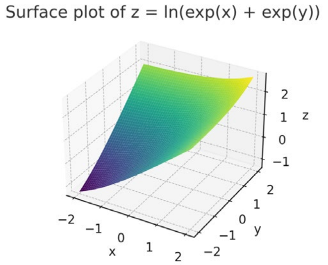
G. Fettweis and L. Thiele, “Algebraic recurrence transformations for massive parallelism,” in IEEE Trans. on Circuits and Systems I (TRANSCC), vol. 40, no. 12, Dec 1993, *DOI:10.1109/81.26903*

Other Examples for Operators ① and ②

①	②
×	+
+	max(a,b)
+	min
min	max
max	min
min	min
+	$\ln\left(e^{a}+e^{b}\right)$

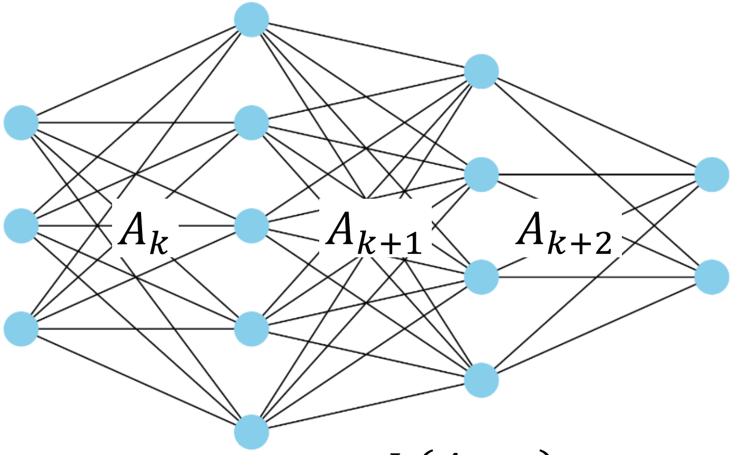
G. Fettweis and L. Thiele, “Algebraic recurrence transformations for massive parallelism,” in IEEE Transactions on Circuits and Systems I, vol. 40, no. 12, Dec 1993

M. Schmidt and G. Fettweis, “On Memory Redundancy in the BCJR Algorithm for Nonrecursive Shift Register Processes,” in IEEE Transactions on Information Theory (TIT), vol. 46, no. 4, Jul 2000



Deep Learning Requirement

Network example



y_k $\qquad y_{k+1} = \Phi(A_k y_k)$

For ML (RELU) need 3rd operator: $y_{k+1} = \Phi \, \textcircled{3} \, A_k y_k$,

$$\textcircled{3} \overset{\Delta}{\rightarrow} RELU(\varphi, x) = \max(\varphi, x)$$

AI/ML – Requires at Least 3 Operators

Generalization for >2 operators holds!

$$a\textcircled{1}[b\textcircled{2}(c\textcircled{3}d)] = a\textcircled{1}[(b\textcircled{2}c)\textcircled{3}(b\textcircled{2}d)]$$

$$= [a\textcircled{1}b \; \textcircled{2} \; a\textcircled{1}c\,] \; \textcircled{3} \; [a\textcircled{1}b \; \textcircled{2} \; a\textcircled{1}\mathrm{d}\,]$$

Iff Operator ① is distributive over ②

Operator ① is distributive over ③

Operator ② is distributive over ③

G. Fettweis and L. Thiele, “Algebraic recurrence transformations for massive parallelism,” in IEEE Transactions on Circuits and Systems I, vol. 40, no. 12, Dec 1993

ML/DL: 3-Operator Matrix Dilemma

$$A = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} \qquad B = \begin{pmatrix} b_{00} & b_{01} \\ b_{10} & b_{11} \end{pmatrix} \qquad C = \begin{pmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{pmatrix}$$

$\odot$ operator 1

$\otimes$ operator 2

$\oplus$ operator 3

$$(A \odot B) \oplus (A \odot C)_{00} = (a_{00} \odot b_{00} \otimes a_{01} \odot b_{10}) \oplus (a_{00} \odot c_{00} \otimes a_{01} \odot c_{10})$$

$$\neq$$

$$A \odot (B \oplus C)_{00} = ((a_{00} \odot b_{00}) \oplus (a_{00} \odot c_{00})) \otimes ((a_{01} \odot b_{10}) \oplus (a_{01} \odot c_{10}))$$

➔ This is where linear Matrix Algebra ends to date

➔ We cannot combine/simplify the operations

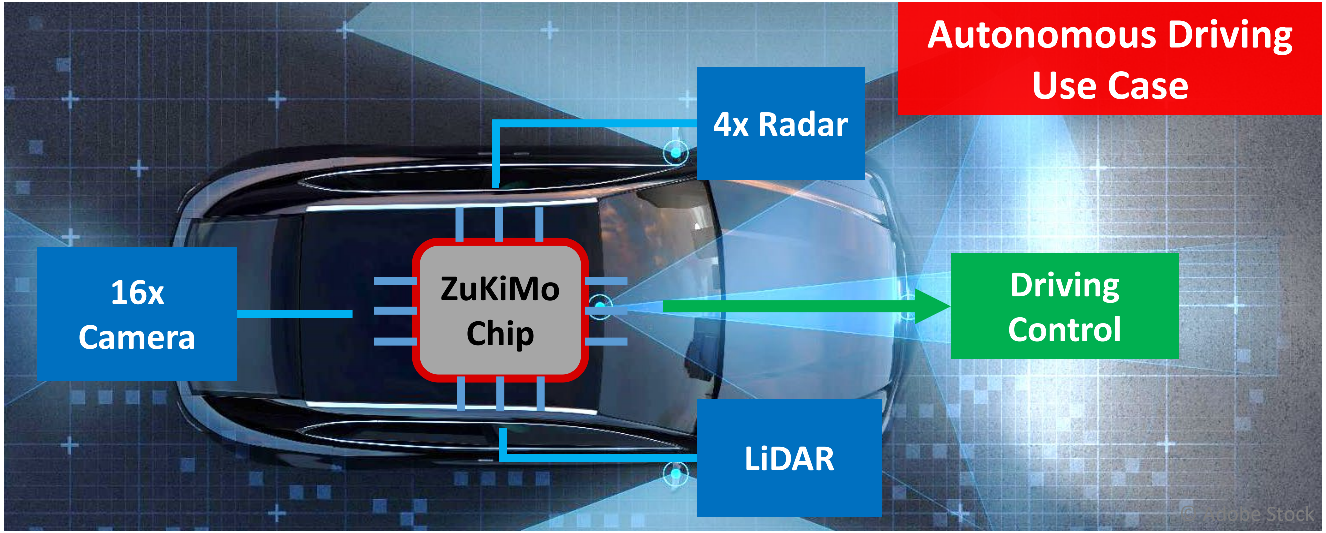
➔ AI requires large matrices to be fetched continuously...

AI/ML – Implementation Challenge

ZuSE KI-mobil - MPSoC with Energy Efficient AI Accelerator

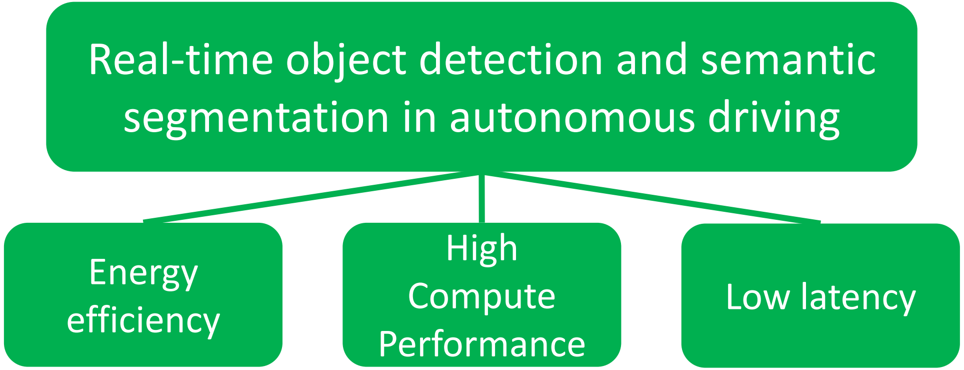
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

- Continuous processing of data from multiple sensors
- Includes **novel AI accelerator** from TU Dresden
- On-chip image pipeline for data pre-processing from Dreamchip



Scientific Coordinator	TECHNISCHE UNIVERSITÄT DRESDEN
Consortium Lead	BMW GROUP
Academia	KIT Karlsruher Institut für Technologie TUM Leibniz Universität Hannover
Industry	Dream Chip Infineon ZF
SME	TECHNOLOGIES TRANSISTORS TRANSFORMATIONS
Technology Partners	arm ARTERIS IP cadence SYNOPSYS GlobalFoundries

Motivation and Overview

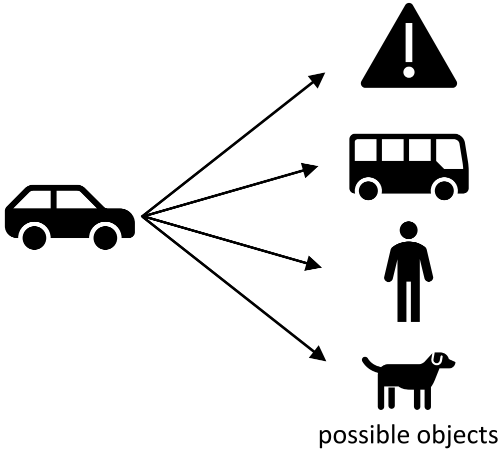


Real-time object detection and semantic segmentation in autonomous driving

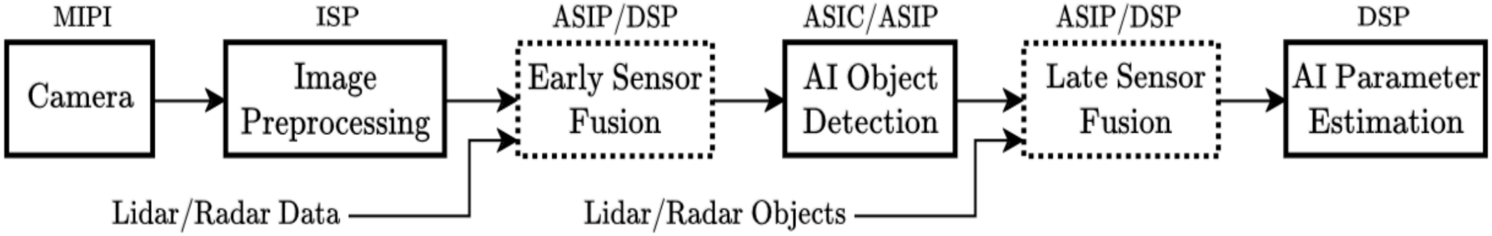
Energy efficiency

High Compute Performance

Low latency



➔ **Image pipeline with heterogenous processing elements required**



MIPI: Camera → ISP: Image Preprocessing → ASIP/DSP: Early Sensor Fusion → ASIC/ASIP: AI Object Detection → ASIP/DSP: Late Sensor Fusion → DSP: AI Parameter Estimation

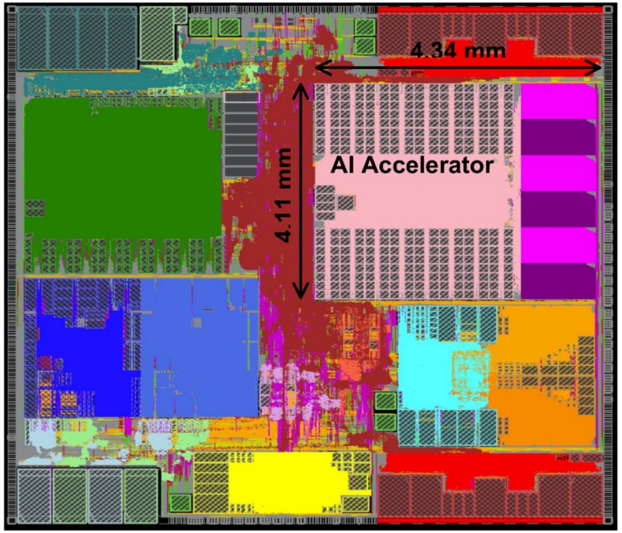
Lidar/Radar Data → Early Sensor Fusion

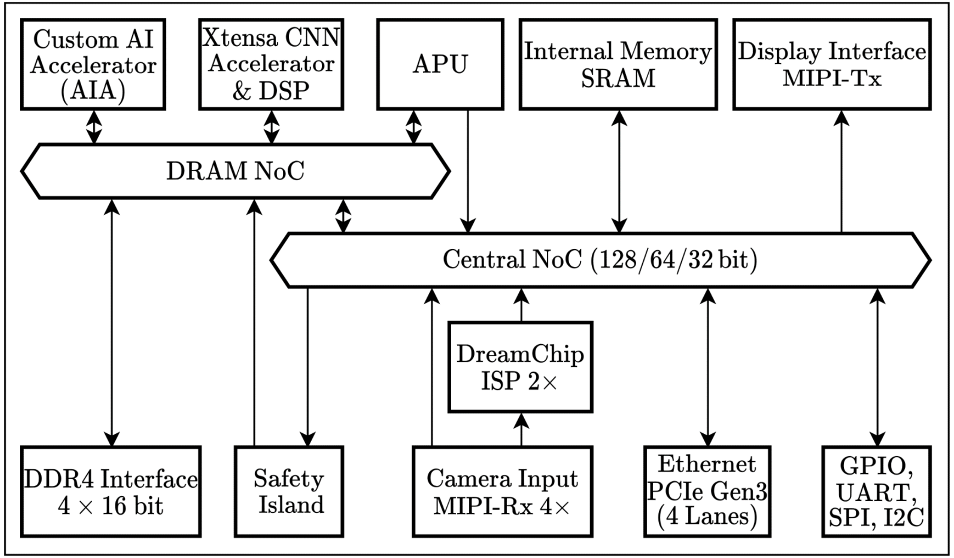
Lidar/Radar Objects → Late Sensor Fusion

Simon Friedrich et al. - COOL Chips 2024

Automotive Platform MPSoC

2023: ZuKiMo ADAS MPSoC
93mm^2 in GF 22FDX, 1.8 billion transistors




Custom AI Accelerator (AIA)
Xtensa CNN Accelerator & DSP
APU
Internal Memory SRAM
Display Interface MIPI-Tx
DRAM NoC
Central NoC (128/64/32 bit)
DreamChip ISP 2×
DDR4 Interface 4 × 16 bit
Safety Island
Camera Input MIPI-Rx 4×
Ethernet PCIe Gen3 (4 Lanes)
GPIO, UART, SPI, I2C

Motivation and Overview

- Overview of SoC:

Compute Units

- APU (Dual-Core ARM Cortex A65AE)
- Safety Island (Dual-Core ARM R52)
- DSP & CNN accelerator (Xtensa NNA110)
- Novel AI accelerator (mixed-precision, D-Conv)

Custom Image Signal Processor (DreamChip ISP)

Internal/External Memory

Peripherals

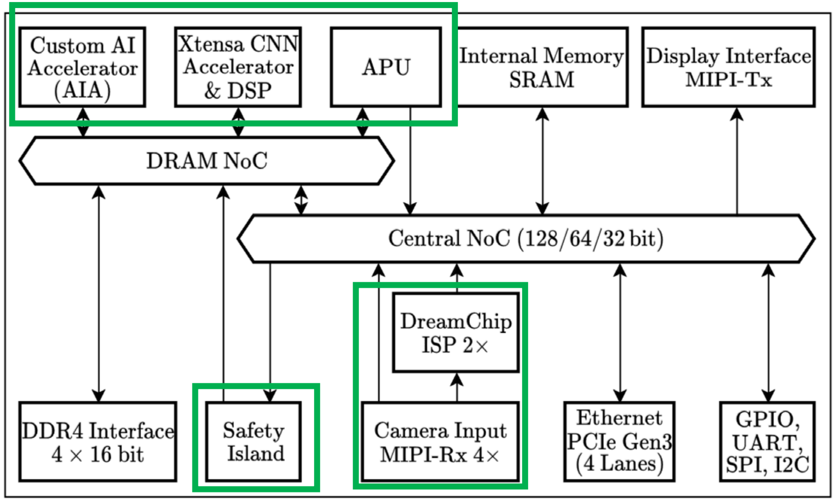
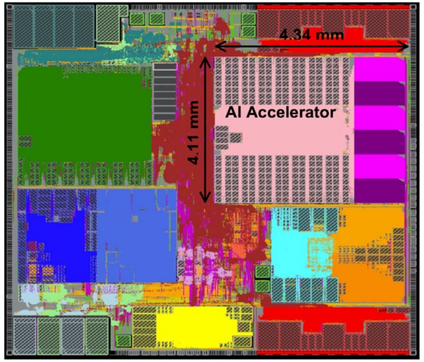

Custom AI Accelerator (AIA)
Xtensa CNN Accelerator & DSP
APU
Internal Memory SRAM
Display Interface MIPI-Tx
DRAM NoC
Central NoC (128/64/32 bit)
DreamChip ISP 2×
DDR4 Interface 4 × 16 bit
Safety Island
Camera Input MIPI-Rx 4×
Ethernet PCIe Gen3 (4 Lanes)
GPIO, UART, SPI, I2C

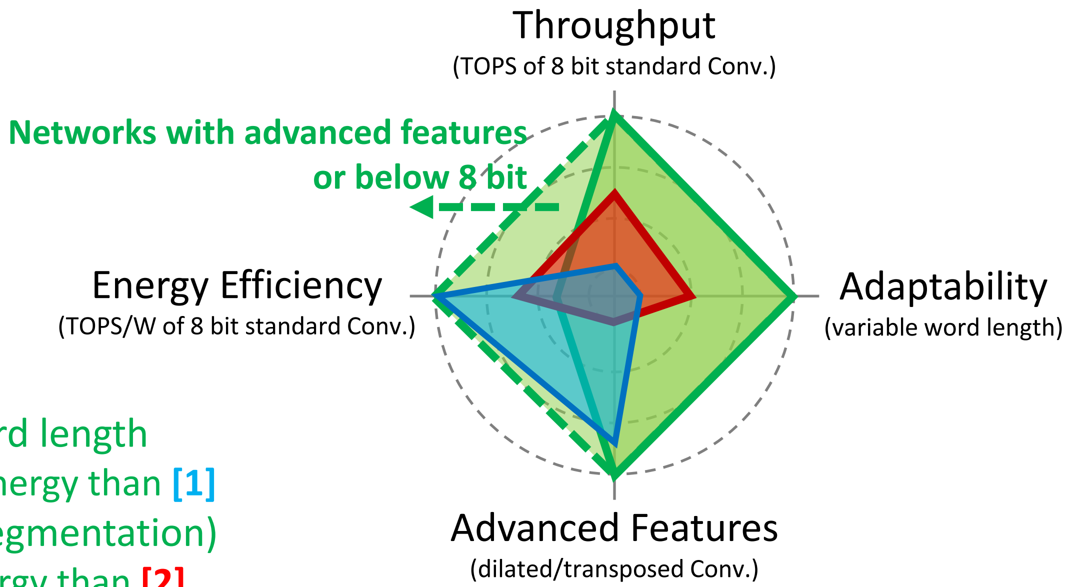
Fig.: Block diagram of SoC supporting image processing pipeline

➔ **Elements of image pipeline integrated into SoC**

➔ **Aim: improve efficiency of AI accelerator for image processing**

ZuKIMo Energy Efficient AI Accelerator: 18mm^2





Networks with layer dependent word length

➔ DeepLabV3+ (4 bit): **2.2x** less energy than **[1]**

Advanced features (e.g. semantic segmentation)

➔ DeepLabV3+ (8 bit): **5x** less energy than **[2]**

Comparing Accelerators:

🟦 Optimized for semantic segmentation [1]
🟥 Optimized for standard convolutions [2]
🟩 ZuKiMo Chip [3]

[1] J. Jung et al., "An Energy-Efficient, Unified CNN Accelerator for Real-Time Multi-Object Semantic Segmentation for Autonomous Vehicle," in IEEE Transactions on Circuits and Systems I: Regular Papers, 2024.
[2] C.-H. Lin et al., "7.1 A 3.4-to-13.3TOPS/W 3.6TOPS Dual-Core Deep-Learning Accelerator for Versatile AI Applications in 7nm 5G Smartphone SoC," in 2020 IEEE International Solid-State Circuits Conference (ISSCC), 2020.
[3] S. Friedrich et al., "A 22 nm 10 TOPS Mixed-Precision Neural Network SoC for Image Processing with Energy-Efficient Dilated Convolution Support," in Proceedings of IEEE Symposium on Low-Power and High-Speed Chips (COOLCHIPS), Apr 2024.

Novel AI Accelerator - Architecture

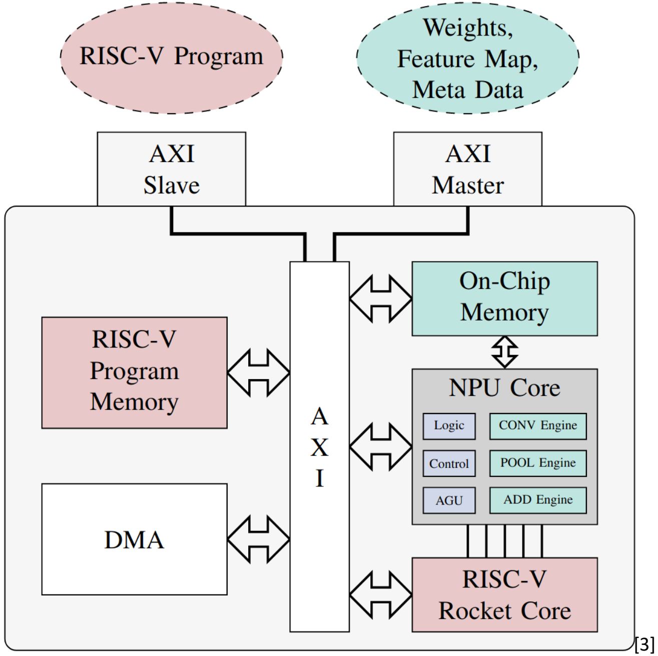


Fig. 2. Block diagram of the accelerator including its external interfaces.

- **Bit-serial Architecture:**
  - Bit-serial processing elements
  - Precision-scalable operands
  - Bit-serial memory design
  - **Performance: 6.144 TOPS (INT8, @1 GHz)**
- **Scalable Architecture:**
  - Variable size of CONV array with broadcasted input data
  - High utilization of processing elements
- **Regular Instruction Set and Memory Mapping:**
  - Lightweight instruction set
  - Efficient acceleration of dilated and transposed convolution

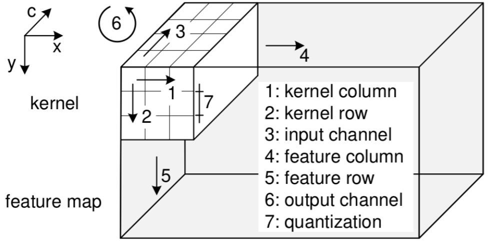
[3] S. Friedrich, S. Balamuthu Sampath, R. Wittig, M. Rohit Vemparala, N. Fasfous, E. Matúš, W. Stechele and G. Fettweis, “Lightweight Instruction Set for Flexible Dilated Convolutions and Mixed-Precision Operands,” in Proceedings of 24th International Symposium on Quality Electronic Design (ISQED 2023), San Francisco, USA, Apr 2023.

Novel AI Accelerator - Architecture

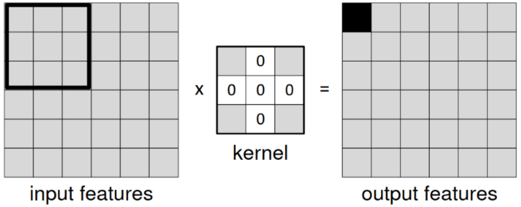
Supported Instructions:

Convolution Layer:

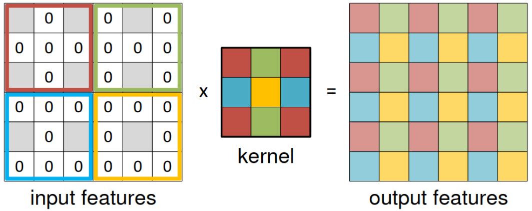
- 1D CONV
- 2D CONV
- Depthwise CONV
- Dilated CONV
- Transposed CONV



(a) Algorithmic dimensions of a convolution layer. Quantization as an additional dimension included.



(b) D-CONV with $R_D = 2$, $K_{x,y} = 2$.



(c) Patterns of a T-CONV with $R_T = 2$, $K_{x,y} = 3$.

[4]

Fully Connected Layer

Padding (Symmetric)

Pooling (max-Pool)

Activation Functions: Custom Function using Piecewise Linear Approximation

Constant Value Addition and Multiplication

Elementwise Addition of 2 Layers (Residual Connection)

S. Friedrich, S. Balamuthu Sampath, R. Wittig, M. Rohit Vemparala, N. Fasfous, E. Matúš, W. Stechele and G. Fettweis, "Lightweight Instruction Set for Flexible Dilated Convolutions and Mixed-Precision Operands," in Proceedings of 24th International Symposium on Quality Electronic Design (ISQED 2023), San Francisco, USA, Apr 2023.

Architecture of Accelerator and Neural Processing Unit

Control Unit (CU):

- RISC-V core, 32 bit, extensions A/C/E
- DNN scheduling, triggers NPU to execute DNN layer
- NPU/DMA configuration
- Issuing data transfers, interrupt handling

Neural Processing Unit (NPU):

- 3 different bit-serial compute engines
- Mixed-precision word length: 2-8 bit
- Zero skipping for dilated and transposed convolution
- Regular addressing scheme:
  - Lightweight instruction set, 60 bit per instruction
  - Footprint 5x smaller than EdgeTPU for MobileNet-v2

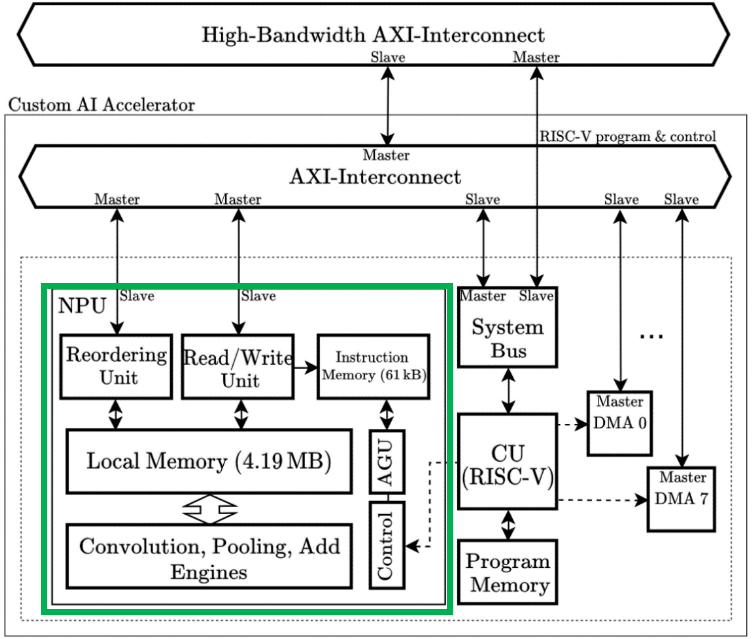


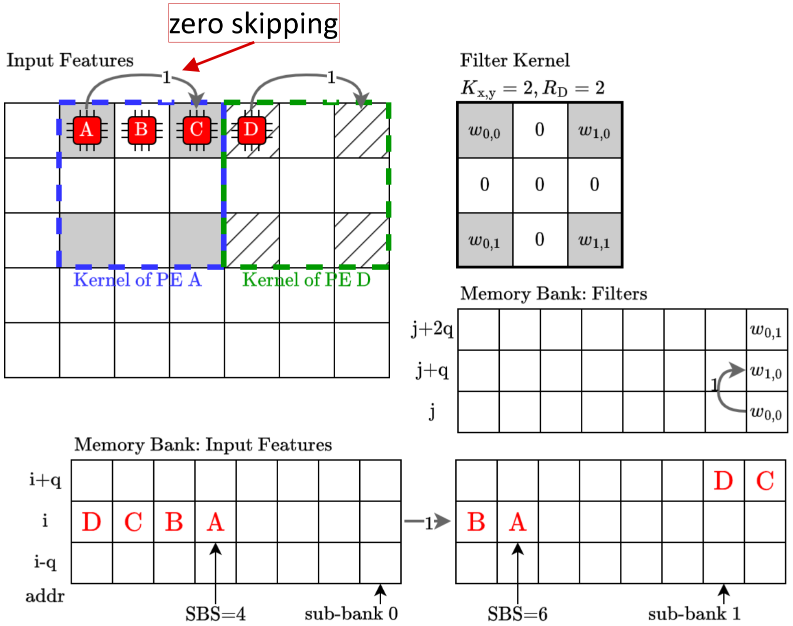
Fig.: AIA including Neural Processing Unit (NPU) and Control Unit (CU)

Efficient Dilated Convolution Support

- <u>Memory Alignment:</u>
  - Unified memory for features, filters, and biases
  - 32 on-chip SRAM banks
  - Each with 8 separately addressable sub-banks
- <u>Address Generation:</u>
  - 32 x 24 OS PEs, 6 partitions with 4 columns each
  - **Initial:** 1 offset, 1 selector per PE column
  - **Runtime:** 1 rel. sub-bank selector (SBS) for all PEs (and 1 rel. bank selector (BS))
  - Strided load unit allows non-adjacent SBS/BS
  - Weights (static) are sorted before runtime, no zeros

➔ **Set SBS to skip adjacent features**

➔ **Only compute operations with non-zero filter values**



$w_{0,0}$	0	$w_{1,0}$
0	0	0
$w_{0,1}$	0	$w_{1,1}$

PE columns (A-D). Non-highlighted features/weights of A and D are not loaded and <u>computations are skipped</u>. Address jump for execution step 1 shown.

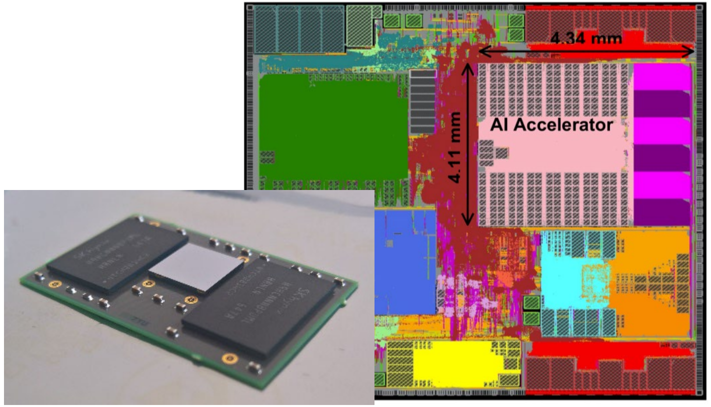
[reference] S. Friedrich et al., “A 22 nm 10 TOPS Mixed-Precision Neural Network SoC for Image Processing with Energy-Efficient Dilated Convolution Support,” in Proceedings of IEEE COOLCHIPS, Apr 2024.

Details of Our AI MPSoC

Technology	FDX 22nm GlobalFoundries™
Area	93.06 mm^2
Logic area	65.14 mm^2
Memory	91.25Mbit @ 27.92 mm^2
Core voltage	0.8V
# transistors	1.8 B
# cameras inputs	16
IP blocks	SYNOPSYS® Dream CHIP
NoC	ARTERIS IP
Cores	cadence arm RISC-V®
MPSoC design	Dream CHIP

6-24 TOPS Programmable AI Accelerator:

- Hardware: TU Dresden – Vodafone Chair
- Compiler: Uni Hanover – IMS
- Chip area: 19%
- On-Chip SRAM: 37%
- Implemented for 1 GHz clock frequency



- Purple: convolution engine (6.144 TOPS)
- Gray: SRAM macros
- Rose: misc. (routing, AGU, CU, DMAs)

Dilated Convolution – Energy Analysis

Our accelerator:

- Efficient support for D-Convs due to zero skipping
- Compute kernels with dimensions $K_{x,y}$

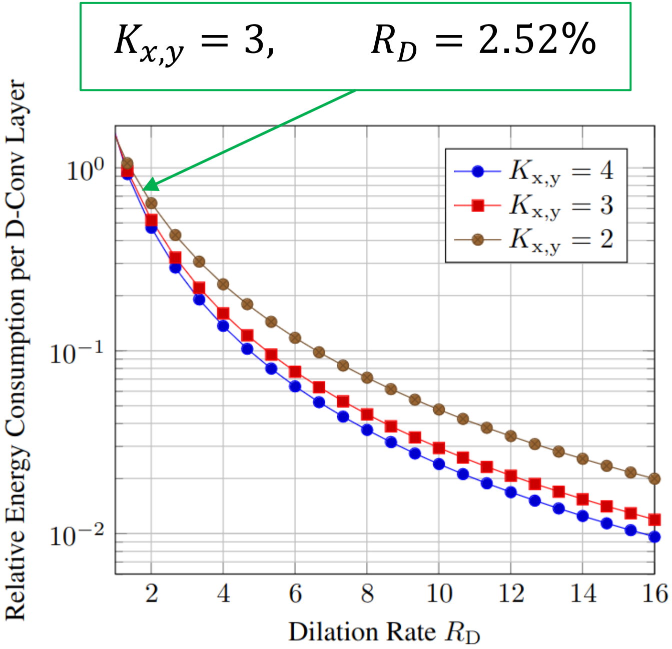
Accelerator ISSCC2020:

- No zero skipping for (dilated) convolutions
- Compute enlarged kernels with dimension $K^{\star}_{x,y}$

$$K^{\star}_{x,y} = \left(K_{x,y} - 1\right) \cdot R_D + 1$$

Results of energy analysis:

- Approximated by energy efficiency and total operations
- Number of computations reduced for $R_D > 1$: reduced energy consumption



$K_{x,y} = 3, \qquad R_D = 2.52\%$

Relative energy consumption per 8 bit D-Conv layer. Compared to ISSCC2020, normalized to 7 nm.

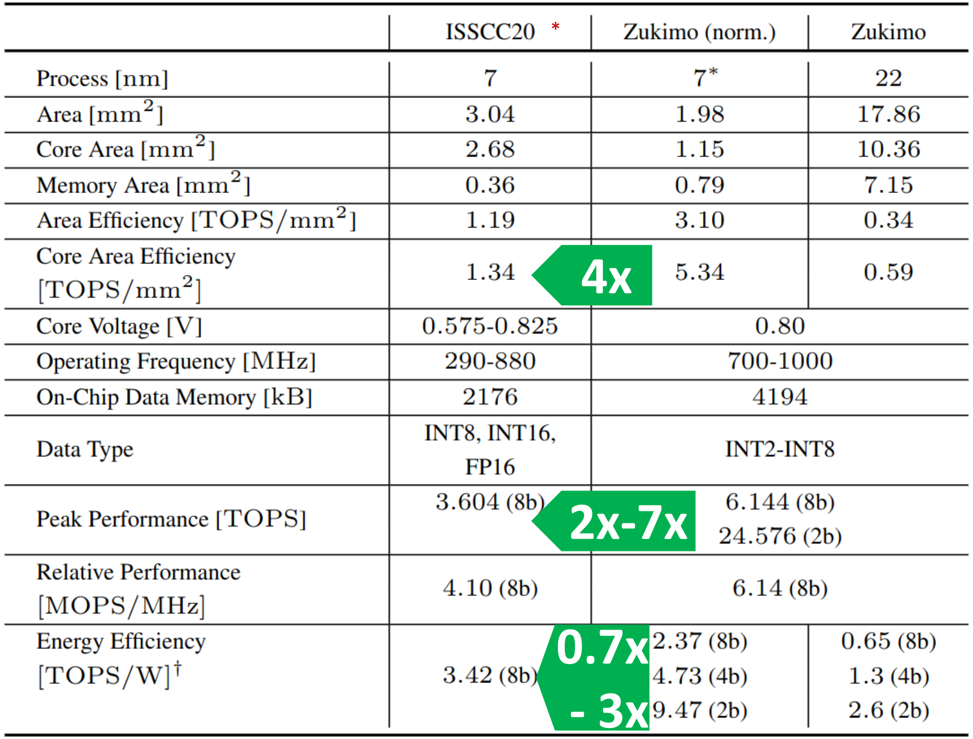
➔ Efficient D-Conv support reduces energy consumption even for DNNs with majority of standard convolution layers

AI Accelerator - Performance Comparison

TOPS/mm²

TOPS

TOPS/W



	ISSCC20 *	Zukimo (norm.)	Zukimo
Process [nm]	7	7*	22
Area [mm^2]	3.04	1.98	17.86
Core Area [mm^2]	2.68	1.15	10.36
Memory Area [mm^2]	0.36	0.79	7.15
Area Efficiency [TOPS/mm^2]	1.19	3.10	0.34
Core Area Efficiency [TOPS/mm^2]	1.34	5.34	0.59
Core Voltage [V]	0.575-0.825	0.80	
Operating Frequency [MHz]	290-880	700-1000	
On-Chip Data Memory [kB]	2176	4194	
Data Type	INT8, INT16, FP16	INT2-INT8	
Peak Performance [TOPS]	3.604 (8b)	6.144 (8b) 24.576 (2b)	
Relative Performance [MOPS/MHz]	4.10 (8b)	6.14 (8b)	
Energy Efficiency [TOPS/W][†]	3.42 (8b)	2.37 (8b) 4.73 (4b) 9.47 (2b)	0.65 (8b) 1.3 (4b) 2.6 (2b)

2023 Zukimo Chip

$$\frac{TOPS}{mm^2\ W} \rightarrow 1.1 \ldots\ 4.2 \times$$

* Normalization from 22 nm to 7 nm technology node. Area reduced by factor 9. Reduction of energy-per-operation by 35 % per node [energy-per-operation].

[†] Our results after 700 MHz synthesis and cycle accurate post-synthesis simulation

* 880 MHz operating frequency

* C. -H. Lin et al., ”7.1 A 3.4-to-13.3TOPS/W 3.6TOPS Dual-Core Deep-Learning Accelerator for Versatile AI Applications in 7nm 5G Smartphone SoC,” in 2020 IEEE International Solid-State Circuits Conference (ISSCC), 2020.

Roadmap for Zukimo Accelerator

Improve current design:

→ Extended quantization following industry trend

Increase of energy efficiency:

→ AVS (adaptive voltage scaling)
→ Optimize compute units
→ Further increase data reusability

Efficient support for Transformers

Additional features