tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

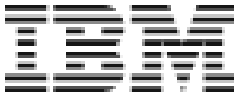
"Empowering the Edge: Advancements in AI Hardware and In-Memory Computing Architectures for TinyML"

Nitin Chawla –Fellow and Director, ST Microelectronics

August 1, 2023

www.tinyML.org

Thank you, **tinyML Strategic Partners**,
for committing to take tinyML to the next Level, together



TINY ML TALKS webcast

Executive Strategic Partners

Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive

Mobile

Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com

TINY ML TALKS webcast

Platinum Strategic Partners

AITRIOS

DEPLOY VISION AI AT THE EDGE AT SCALE

SONY

TINY ML TALKS webcast

Gold Strategic Partners

ANALOG DEVICES

AHEAD OF WHAT'S POSSIBLE™

ANALOG DEVICES

AHEAD OF WHAT'S POSSIBLE™

Where what if becomes what is.

Witness potential made possible at analog.com.


Build the
Future of tinyML
on arm

TINY ML TALKS webcast

EDGE IMPULSE

The Leading Development Platform for Edge ML

edgeimpulse.com

Driving decarbonization and digitalization. Together.

Infineon serving all target markets as
Leader in Power Systems and IoT

www.infineon.com

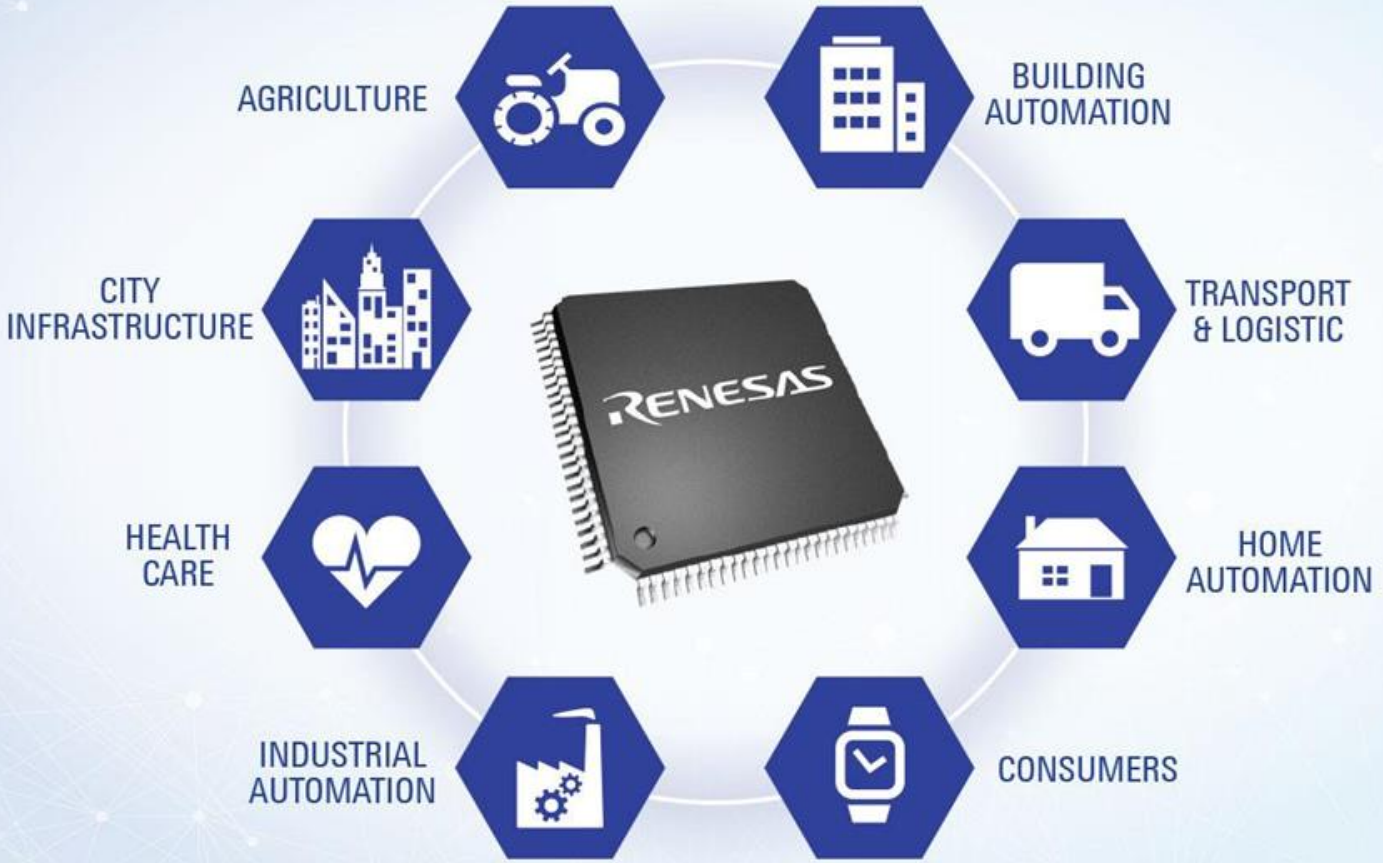
innatera

NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE

www.innatera.com


Microsoft

Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.



renesas.com

STMicroelectronics provides extensive solutions to make tiny Machine Learning easy

life.augmented

www.st.com/ai

synaptics®

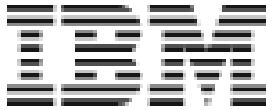
ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com

TINY ML TALKS webcast

Silver Strategic Partners



Join Growing tinyML Communities:

16k members in
49 Groups in 41 Countries

tinyML - Enabling ultra-low Power ML at the Edge

https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/

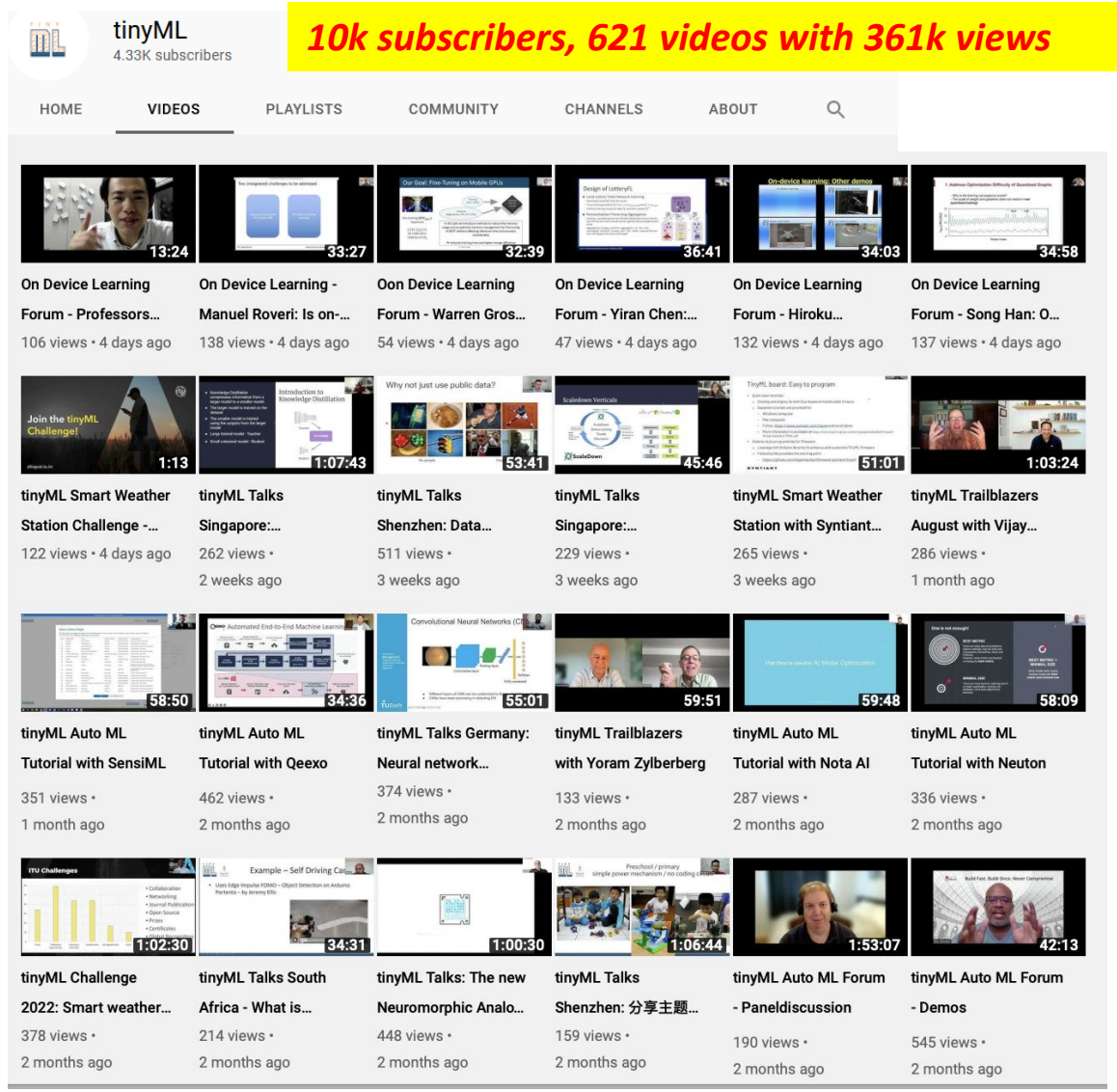
4k members
&
12.7k followers

The tinyML Community

https://www.linkedin.com/groups/13694488/

TINY ML TALKS webcast

YouTube



Subscribe to tinyML YouTube Channel for updates and notifications

(including this video)

www.youtube.com/tinyML

TINY ML FOUNDATION

Call for Presentations and Posters – Deadline August 7
https://www.tinyml.org/event/asia-2023/

WEVOLVER

https://www.wevolver.com/article/2023-edge-ai-technology-report

TINY ML FOUNDATION

2023 Edge AI Technology Report

The guide to understanding the state of the art in hardware & software in Edge AI.

TINY ML TALKS webcast

Nitin Chawla

Nitin Chawla is an ST Fellow and Director in the Technology R&D (Strategy and Innovation) organization at STMicroelectronics, where he leads the research initiatives in the area of Low Power Neural Networks and In-Memory computing architectures for Edge and Tiny ML applications. Nitin has a major in Electronic Circuits and Systems. He is an alumnus of Stanford University and holds a TRIZ diploma from the Massachusetts Institute Of Technology, Cambridge. He has served in different R&D and product organizations over the last 25 years. Before joining STMicroelectronics, he was the Chief Scientist of the HLS Product Division at Mentor Graphics Corporation based in Oregon, USA. Nitin has over 40+ US patents and more than 30 conference and journal publications.

life.augmented

Case Study: Digital SRAM In-Memory Computing Multi-Tiled Neural Processing Unit for Ultra Low Power Inference Applications

Nitin Chawla, Giuseppe Desoli and the ST “Orlando” Team:

Outline

- Introduction
- In Memory NPU architecture
- SRAM DIMC tile
- Silicon results
- Mapping strategies
- Inference examples
- Conclusions

life.augmented

Outline

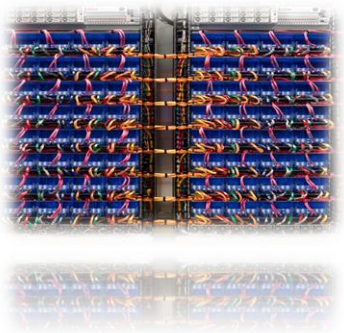
- **Introduction**
- In Memory NPU architecture
- SRAM DIMC tile
- Silicon results
- Mapping strategies
- Inference examples
- Conclusions

life.augmented

AI Applications from Cloud to Tiny Machine Learning

Back-end Training (1x)

- Big training data
- Big models
- Fast iteration

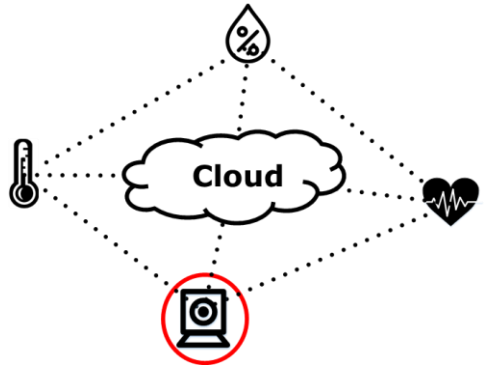


Online/Cloud AI Processing (100x)

- Cloud AI ASIC
- Analysis & recognition
- Power demands

Intelligent Edge Devices (100,000x)

- SoC with NPU accelerators
- Optimized algorithms and CNN-light



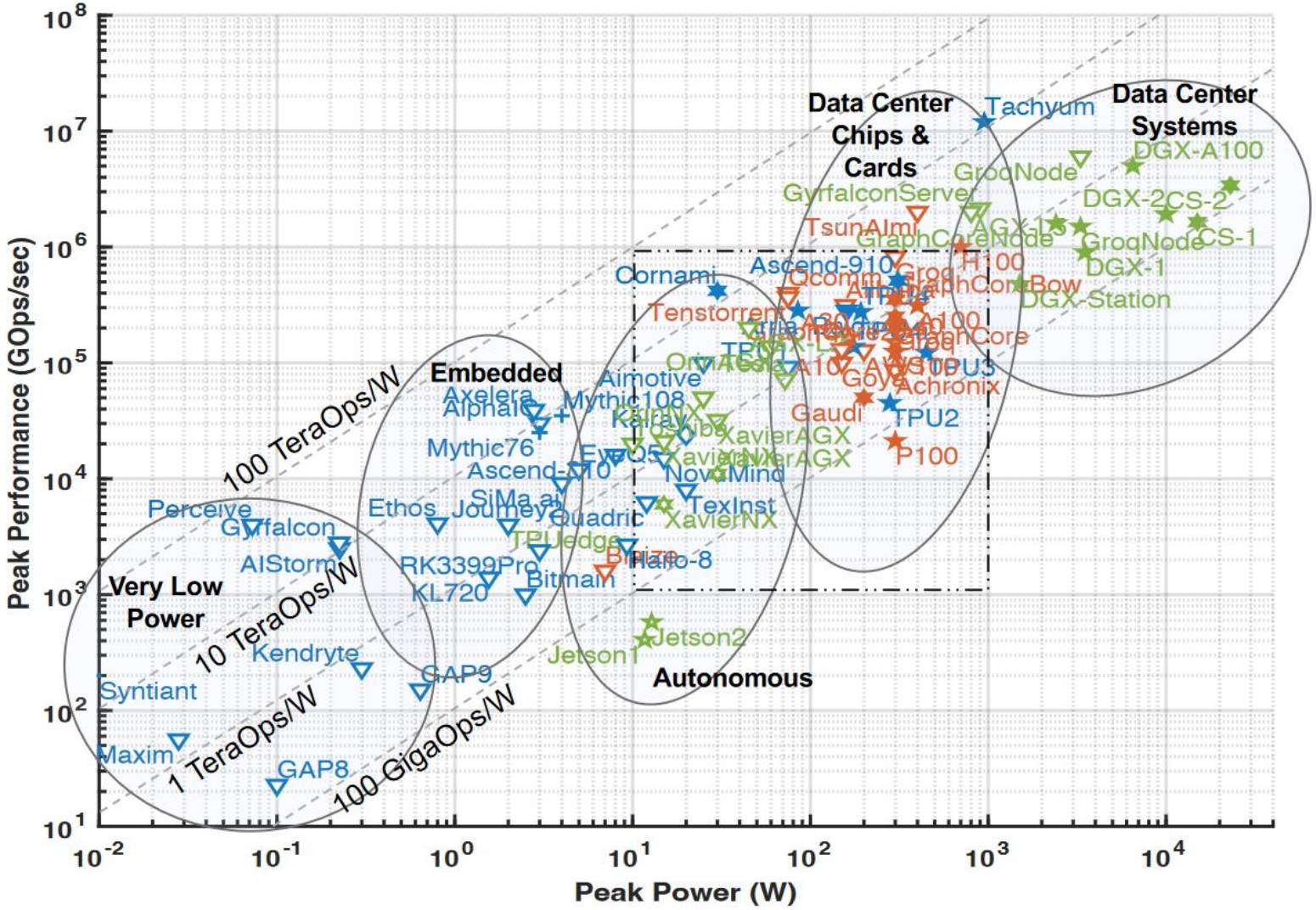
Intelligent Tiny Devices (1,000,000x)

- MCU with HW accelerators
- Very tiny models & computation

CNN: Convolutional Neural Networks
NPU: Neural Processing Unit
MCU: Micro Controller Unit

life.augmented

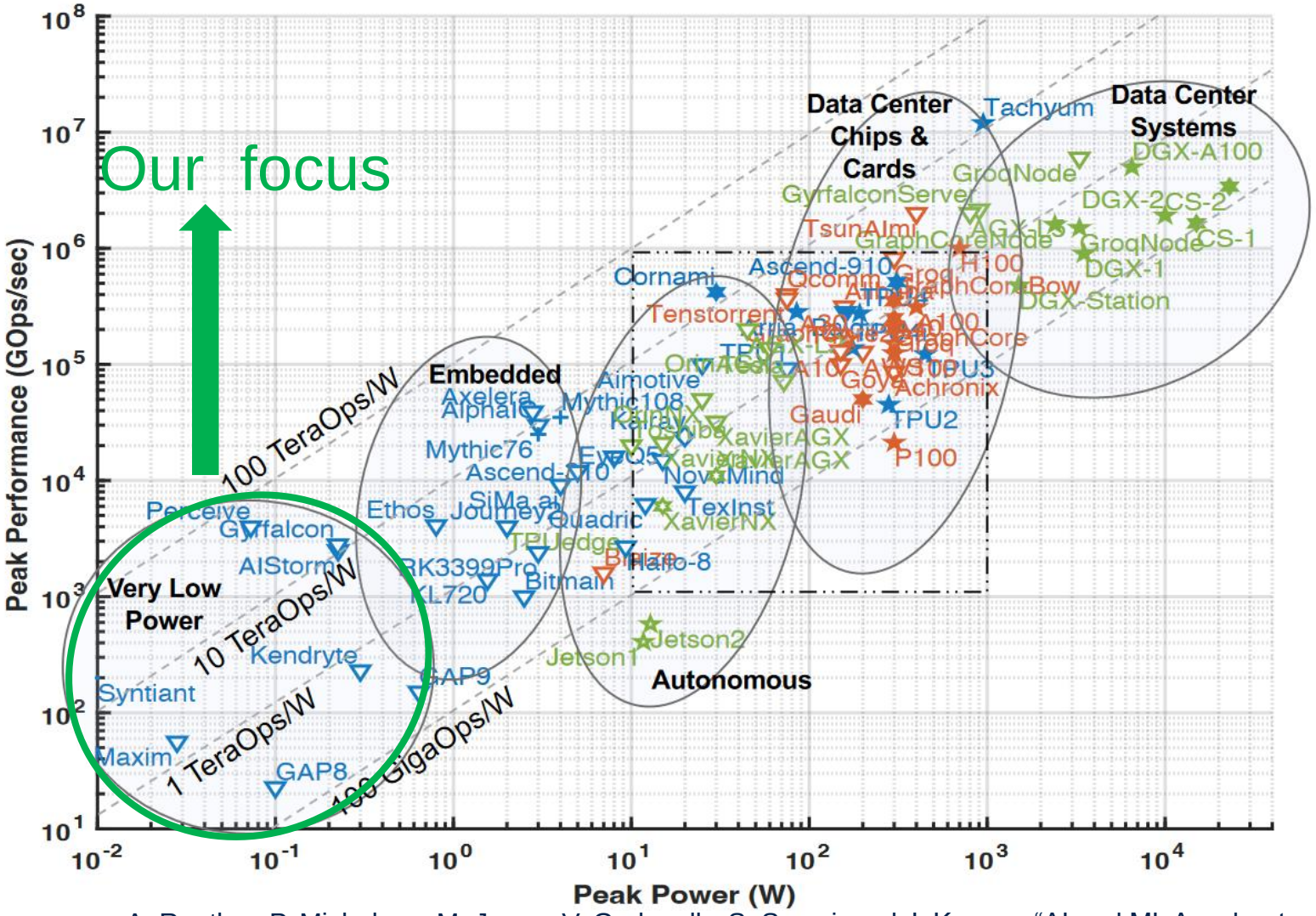
Deep Learning architecture: a Large space



Name	Type
Arm Ethos U-65	Dataflow
GreenWaves GAP 89	Dataflow
Intel Movidius Myriad X	Hybrid RISC-DSP-GPU
Mythic	Analog Flash IMC
Arm Ethos N78	Dataflow
Kneron KL720	Dataflow
Nvidia Jetson Orin	GPU
Renesas DRP-AI	Dataflow / reconfigurable
NXP Neutron eIQ	Dataflow

A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "AI and ML Accelerator Survey and Trends," in *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, Sep. 2022

Deep Learning Architecture: a Large Space

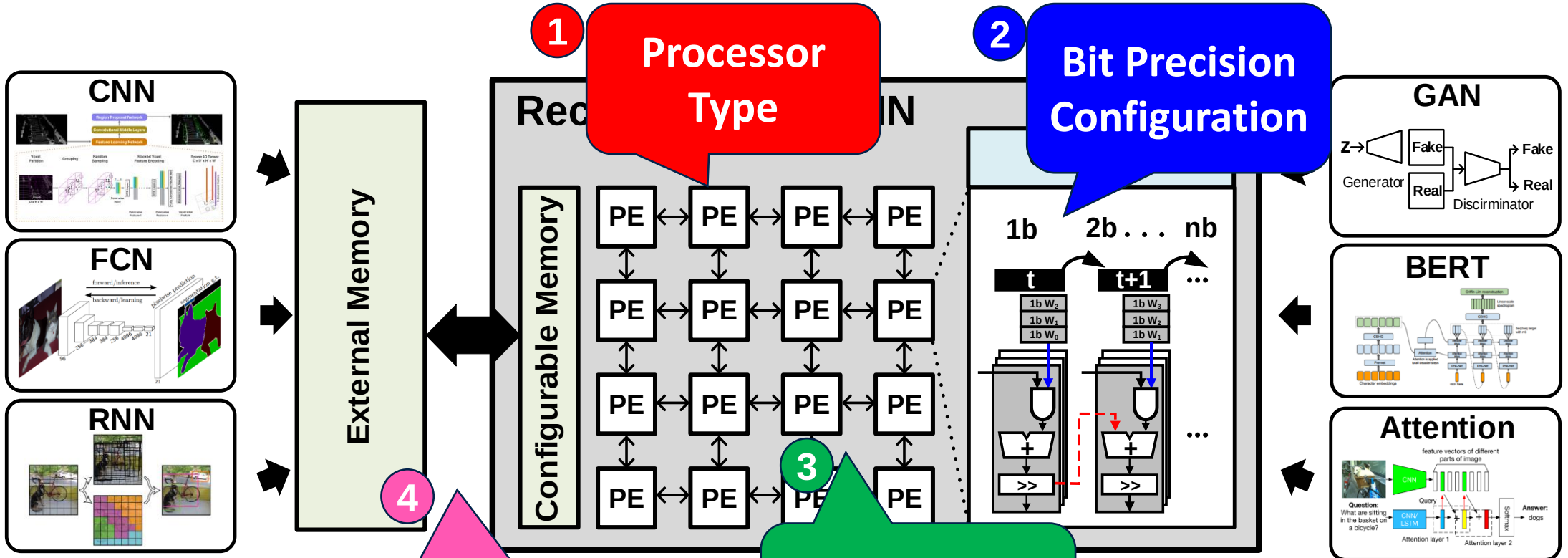


Name	Type
Arm Ethos U-65	Dataflow
GreenWaves GAP 89	Dataflow
Intel Movidius Myriad X	Hybrid RISC-DSP-GPU
Mythic	Analog Flash IMC
Arm Ethos N78	Dataflow
Kneron KL720	Dataflow
Nvidia Jetson Orin	GPU
Renesas DRP-AI	Dataflow / reconfigurable
NXP Neutron eIQ	Dataflow

A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "AI and ML Accelerator Survey and Trends," in *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, Sep. 2022

Key factors for Deep Learning Hardware

- Compute Density and Energy Efficiency are key Figure of Merits(FOM)

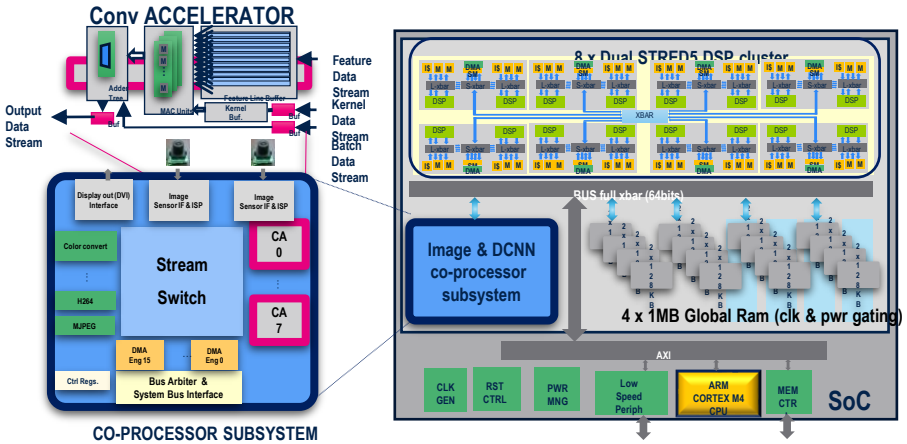


CNN: Convolutional Neural Networks
FCN: Fully Convolutional Networks
RNN: Recurrent Neural Networks

GAN: Generative Adversarial Networks
BERT: Bidirectional Encoder Representation from Transformers

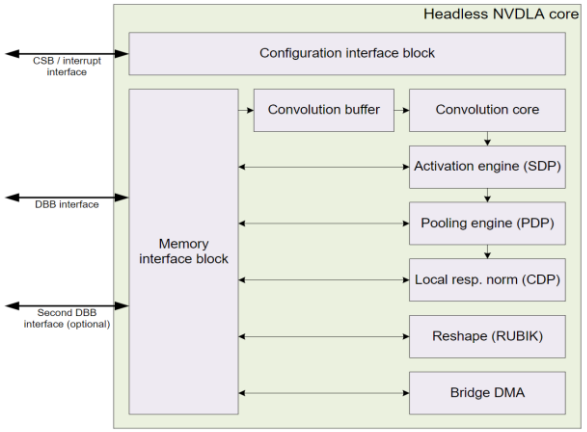
life.augmented

Examples of Embedded AI architectures



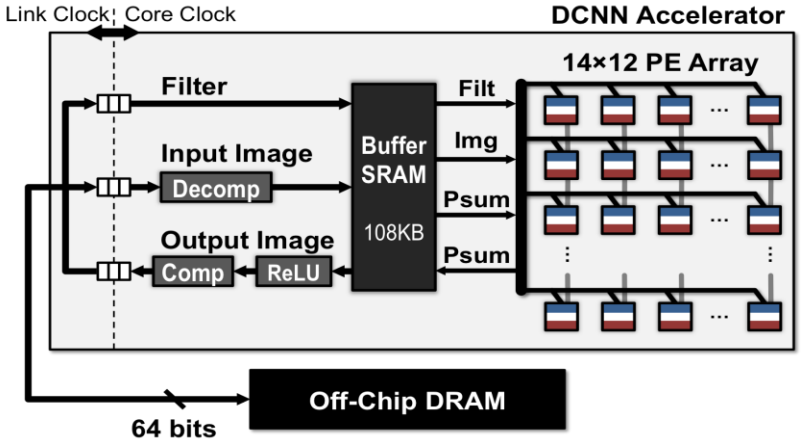
2017

A 2.9 TOPS/W deep convolutional neural network SoC in FD-SOI 28nm for intelligent embedded systems," ISSCC 2017 https://doi.org/10.1109/ISSCC.2017.7870349



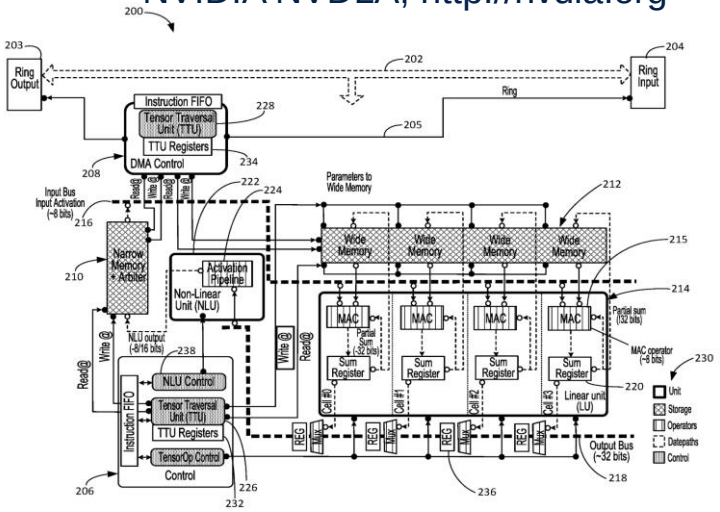
2018

NVIDIA NVDLA, http://nvdla.org



2016

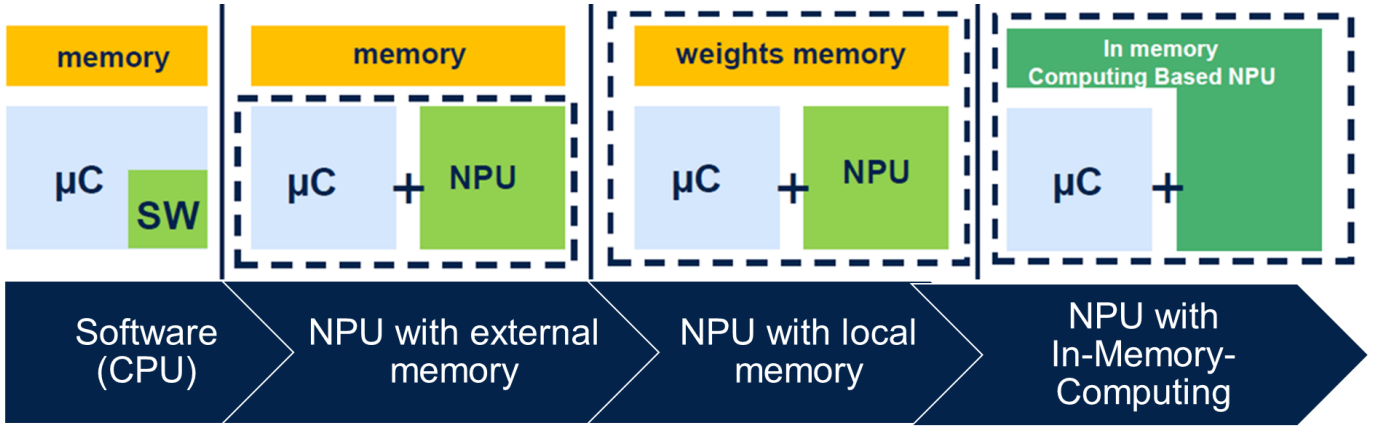
MIT Eyeriss http://eyeriss.mit.edu/



2019

Google Coral edge TPU US20190050717A1

NPU roadmap: towards In Memory compute



Improved Compute Density and Energy Efficiency

In-memory compute taxonomy

Analog IMC

- Approximate BL accumulation
- Bit cell Vt variation limits row parallelism
- Readout throughput limited by ADC
- Approximate compute with complex BIST/Functional test screening

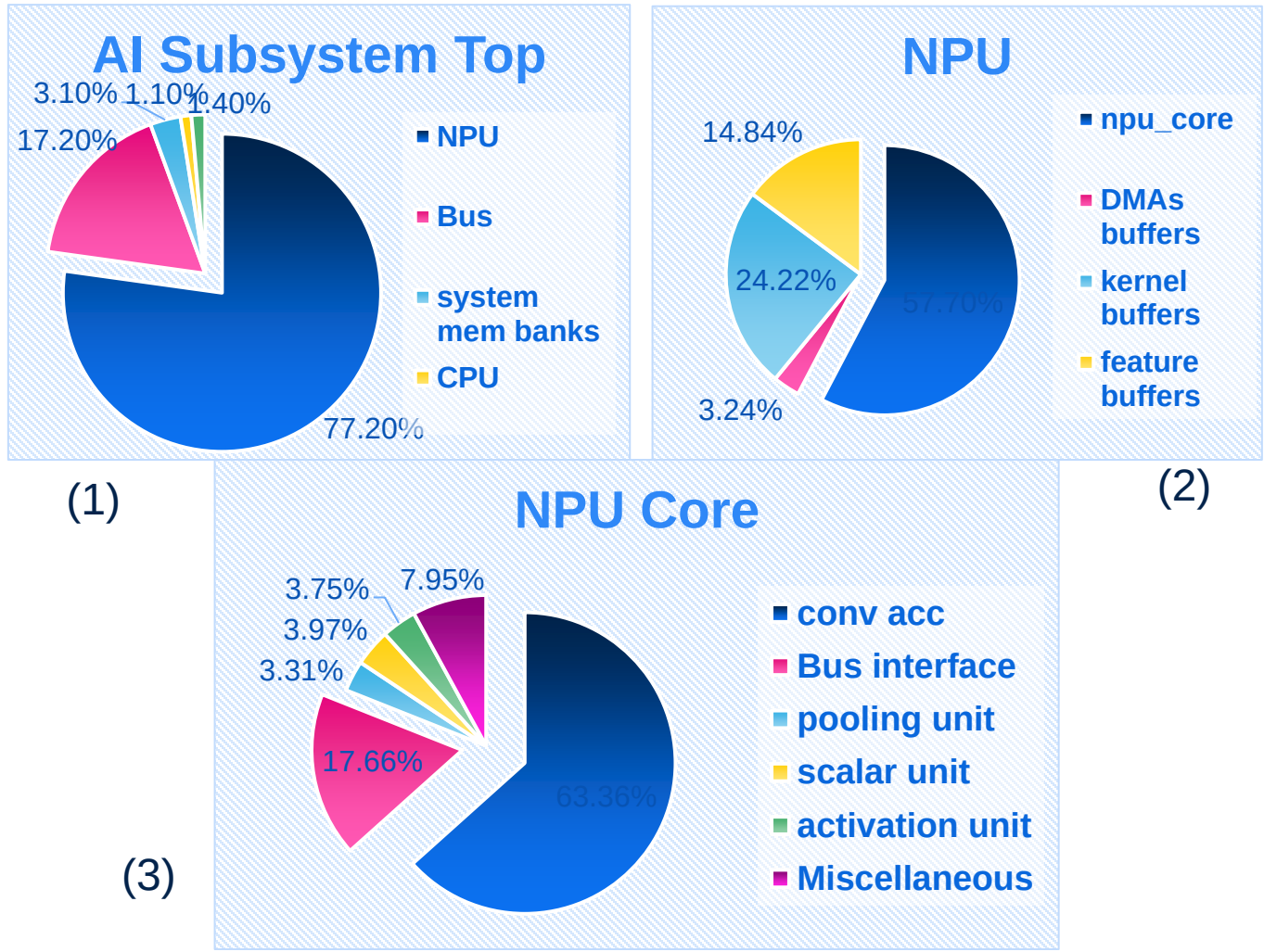
Digital IMC

- Deterministic and dataflow compatible
- Pushed rule/Logic Bitcell
- Duality of **memory** and **computational** modes
- Wide support of DVFS and Adaptive Body Bias
- Deterministic compute for DFT & Safety needs

life.augmented

NPU power consumption vs compute density

- Key FOM: **TOPS/W & TOPS/mm2**
- These metrics vary based on:
  - **NPU architectural style**
    data-flow, bit precision, sparsity support, weights compression
  - **Operating modes**
    input/weights/output stationary
  - **Process technology**
    40nm to 12nm for typical edge@AI Socs
- **1 and 5 TOPS/W** for current MCUs process technology options
- **50-200 TOPS/W expected to be needed in the next 5 years for AI@edge**



Relative Power consumption for a typical NPU:
(1) System Level, (2) NPU + Mem, (3) NPU Core

life.augmented

Outline

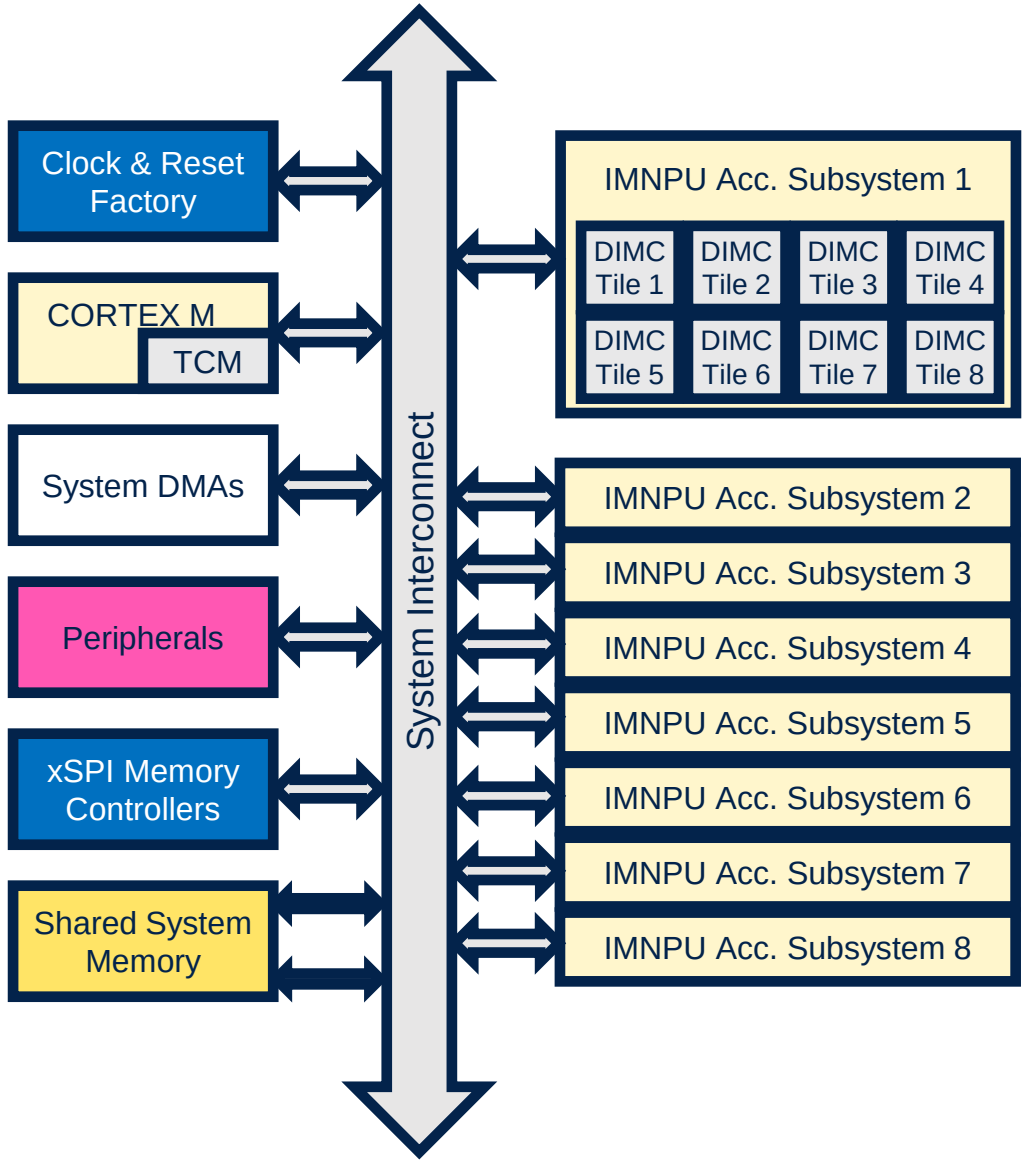
- Introduction
- **In Memory NPU architecture**
- SRAM DIMC tile
- Silicon results
- Mapping strategies
- Inference examples
- Conclusions

life.augmented

Prototype SOC Architecture

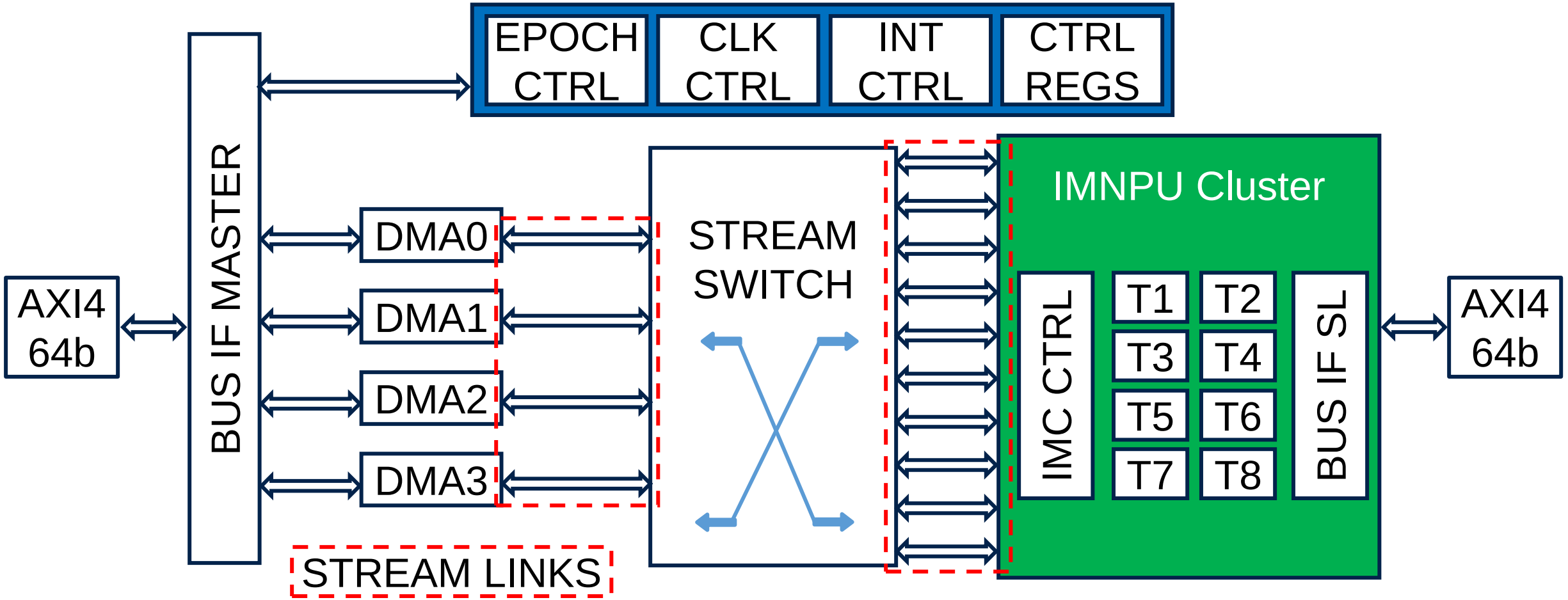
System Components

- Cortex M Host
- 8 IMNPU Subsystems
- Peripherals to load Inputs
- External memory controllers
- Shared system memories for weight storage
- System Interconnect



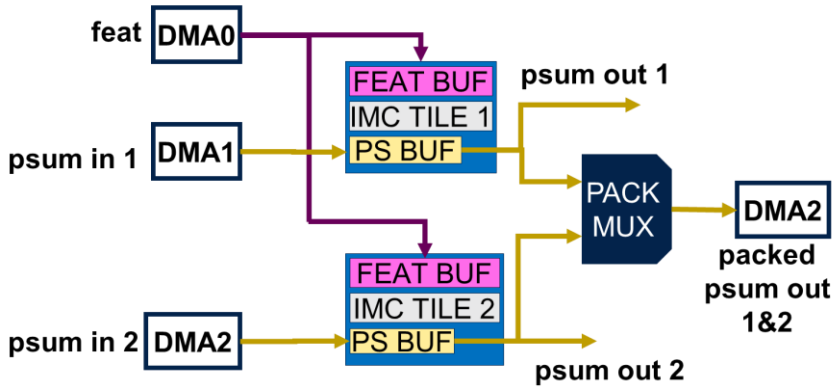
life.augmented

IMNPU accelerator subsystem



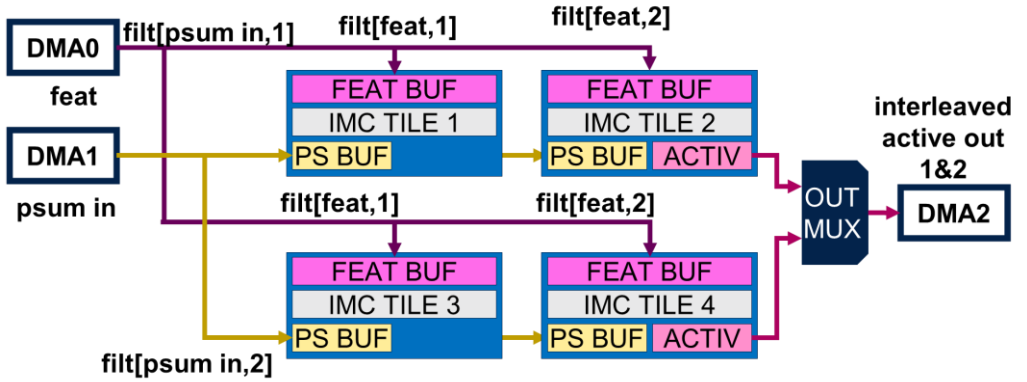
Architectures support for chaining and tiling

2 tiles with PACK MUX



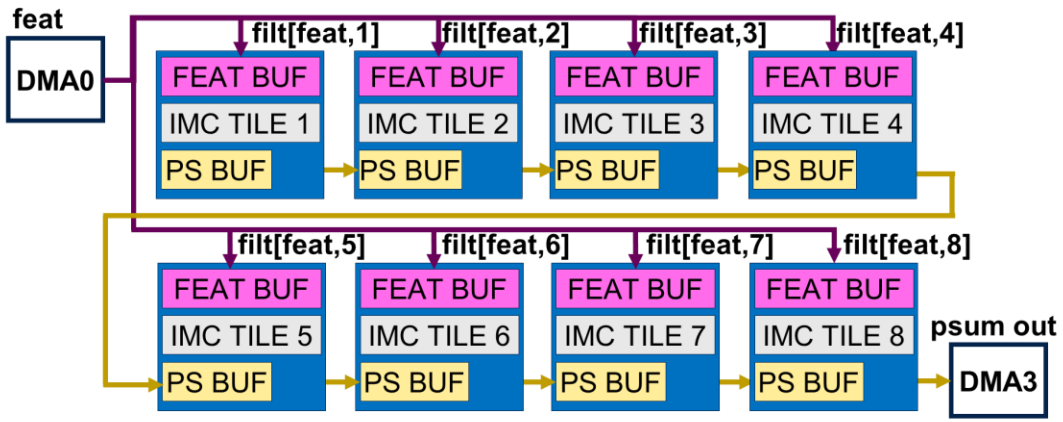
Two parallel IMC tiles generating packed and unpacked partial sum output data

2 chains of 2 tiles with OUT MUX



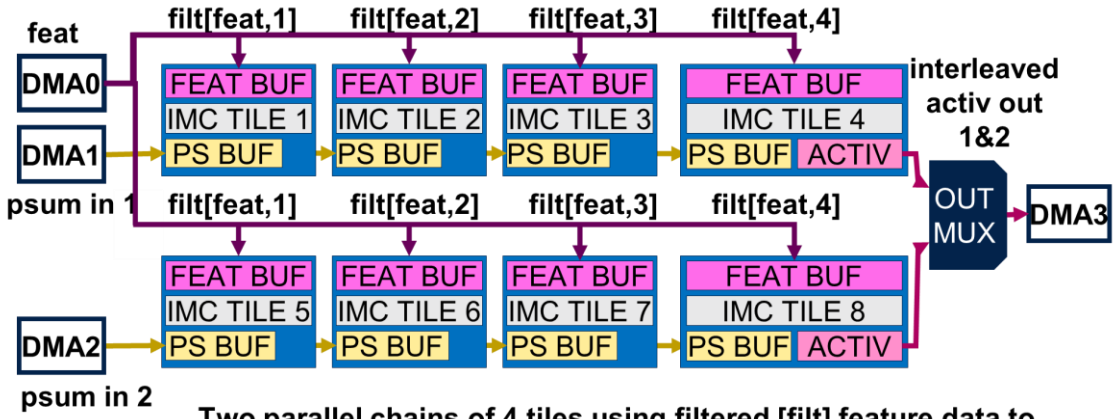
Two parallel chains of 2 tiles using packed and filtered [filt] psum to generate interleaved activation output data

All tiles chained together



One chain of 8 IMC tiles using filtered [filt] feature data without partial sum in data to generate partial sum output data

2 chains of 4 tiles with OUT MUX



Two parallel chains of 4 tiles using filtered [filt] feature data to generate interleaved activation output data

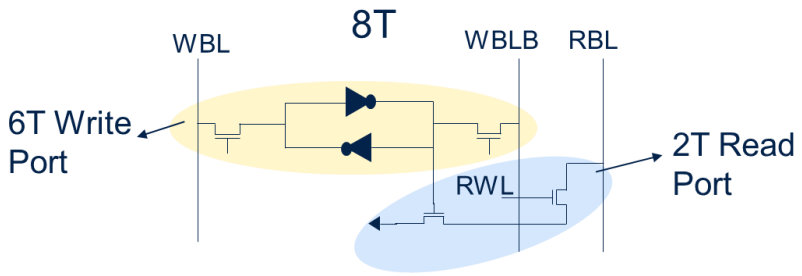
life.augmented

Outline

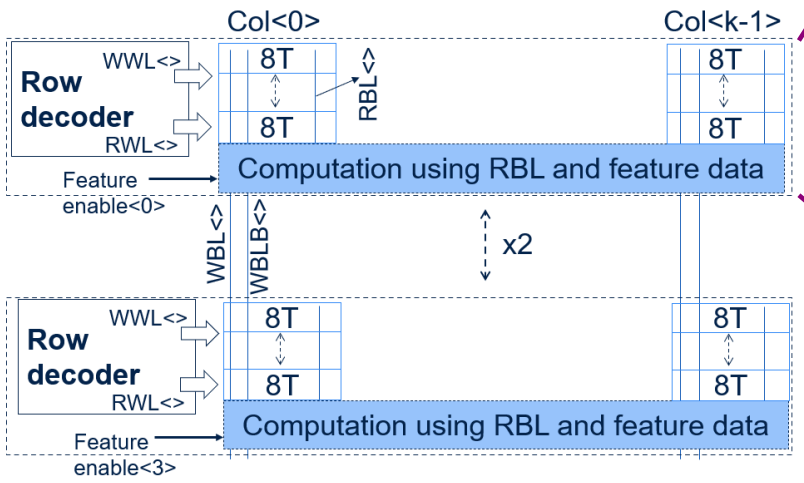
- Introduction
- In Memory NPU architecture
- **SRAM DIMC tile**
- Silicon results
- Mapping strategies
- Inference examples
- Conclusions

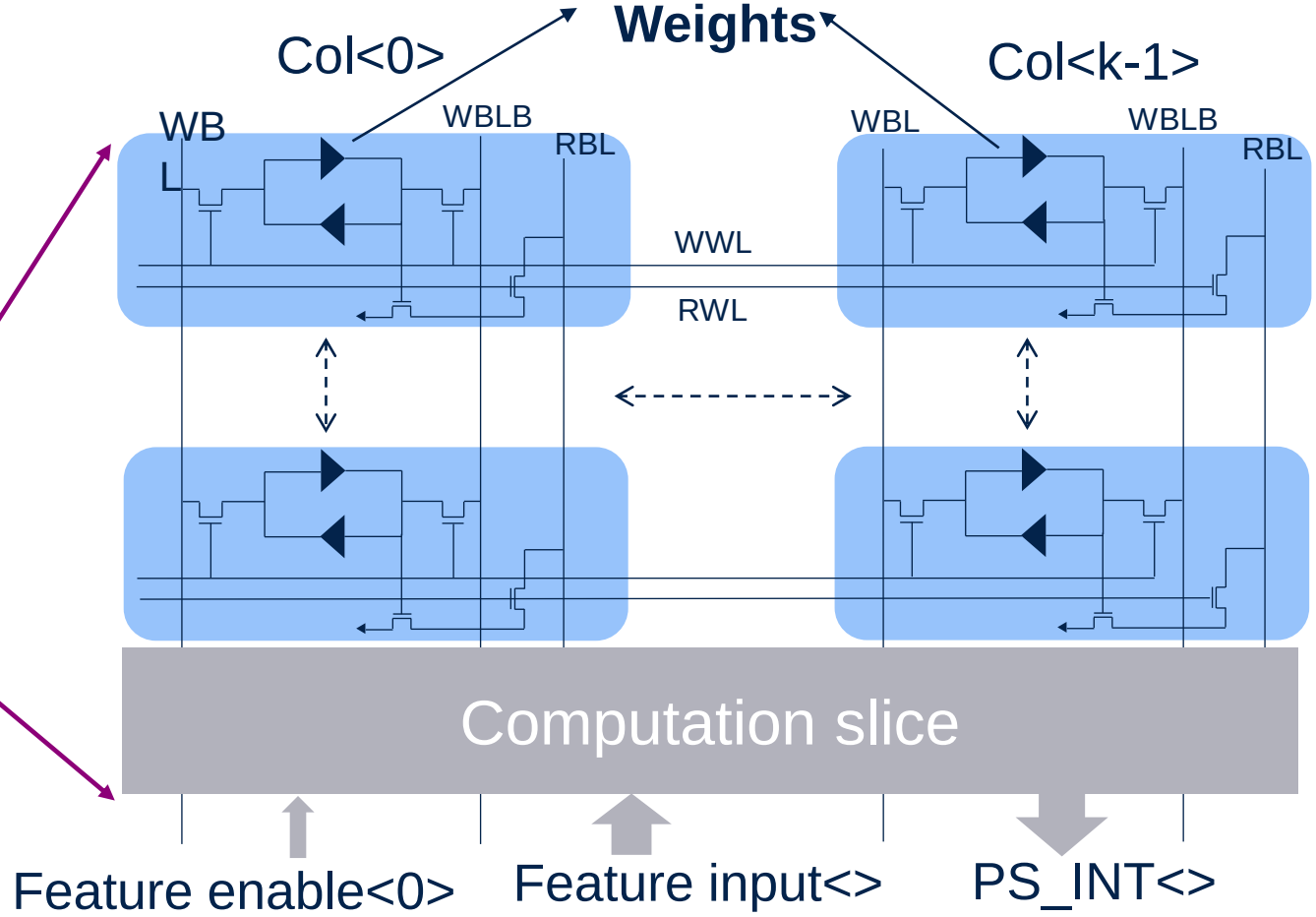
life.augmented

SRAM DIMC array segment arrangement

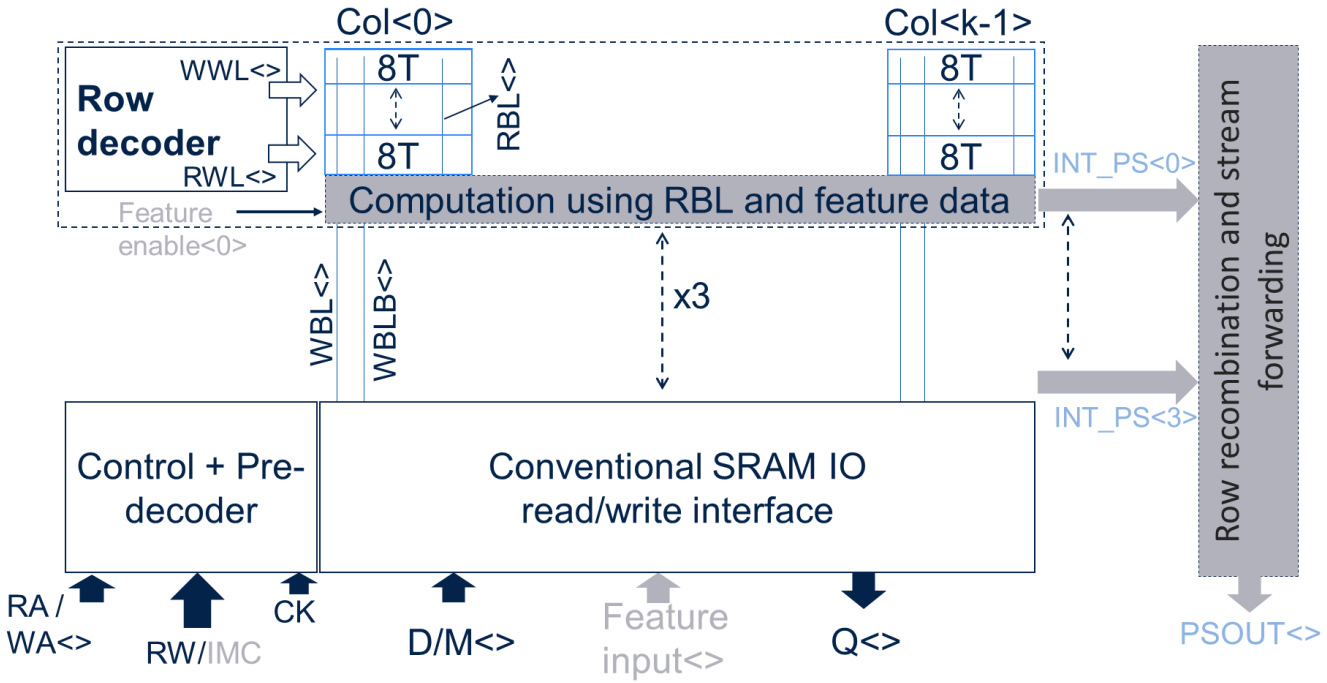


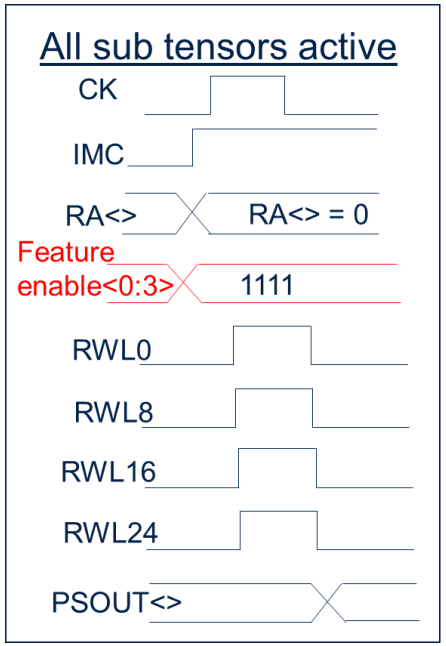
- 1R1W 8T bitcell based core array
- Read port decoupled from Write and provides good performance scaling, benefitting from body bias(BB) strategy of FD-SOI
- Better Vmin vs conventional 6T based SRAM

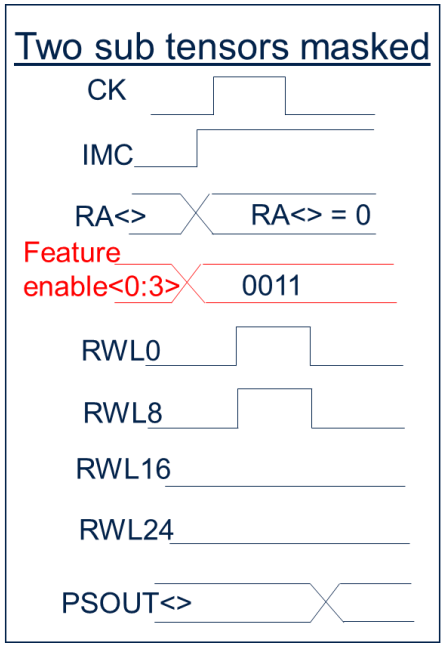




DIMC Instance architecture







- Address decode scheme enabling 4 segments parallel access (e.g. 32 rows)
- Computation supports full tensor/sub tensor modes
- Unused tensor space can be gated

life.augmented

Outline

- Introduction
- In Memory NPU architecture
- SRAM DIMC tile
- **Silicon results**
- Mapping strategies
- Inference examples
- Conclusions

life.augmented

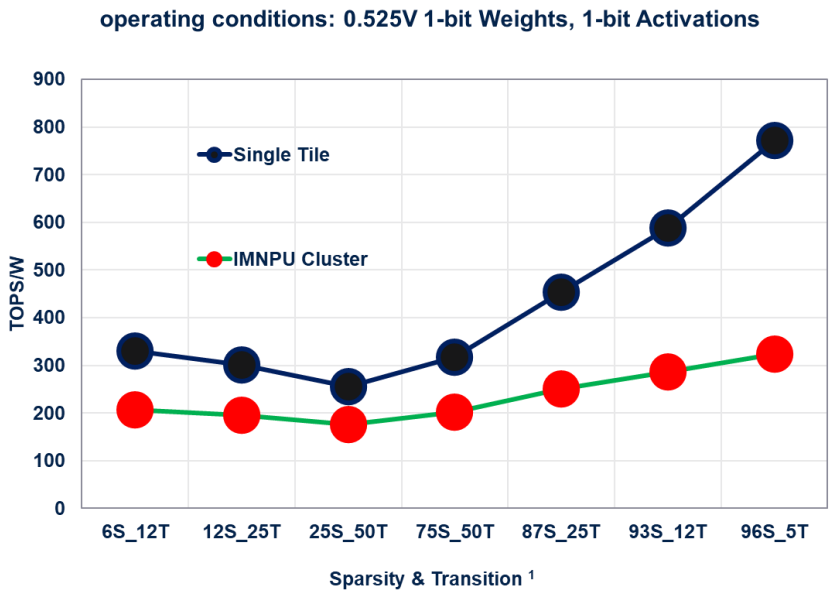
In Memory NPU silicon measurements

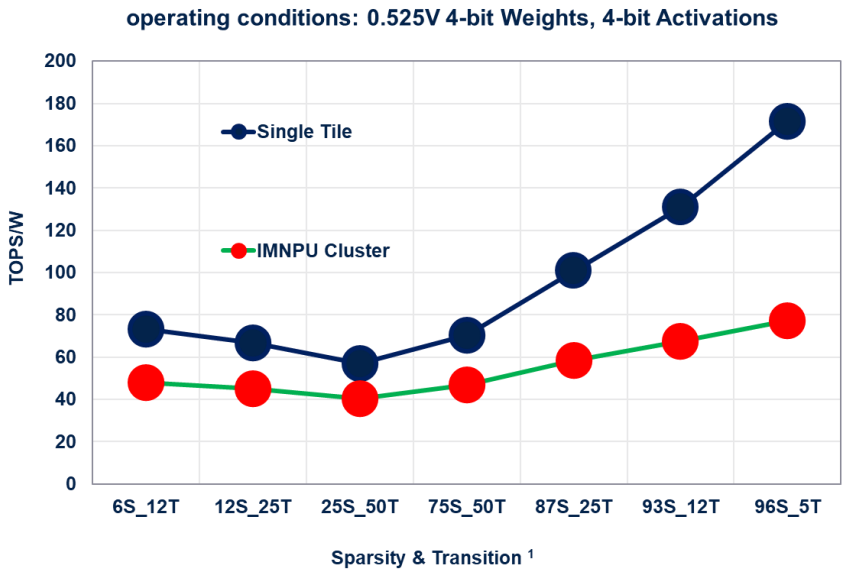
TOPS/W (4bW-4bF): Tile: **176 TOPS/W, IMNPU: 76 TOPS/W**

TOPS/W (1bW-1bF): Tile: **770 TOPS/W, IMNPU: 310 TOPS/W**

Better energy efficiency for sparse networks

Tile level energy performance gains diminish with data movement costs

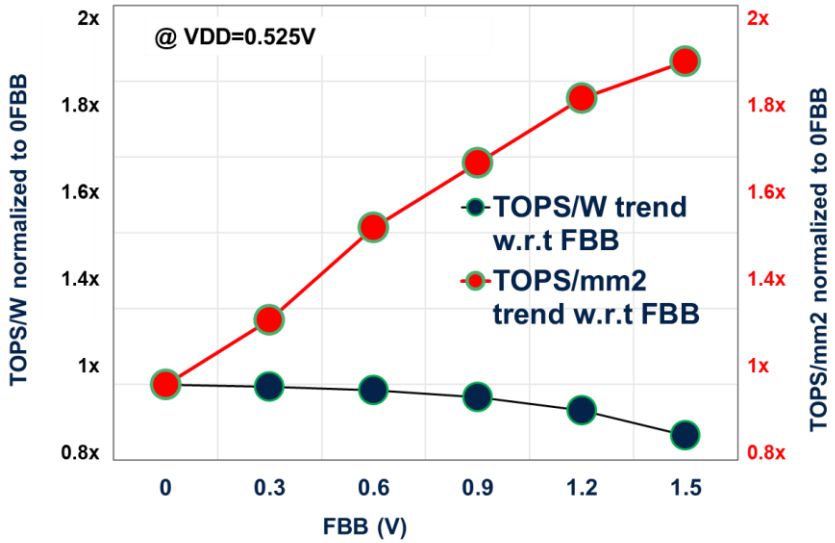


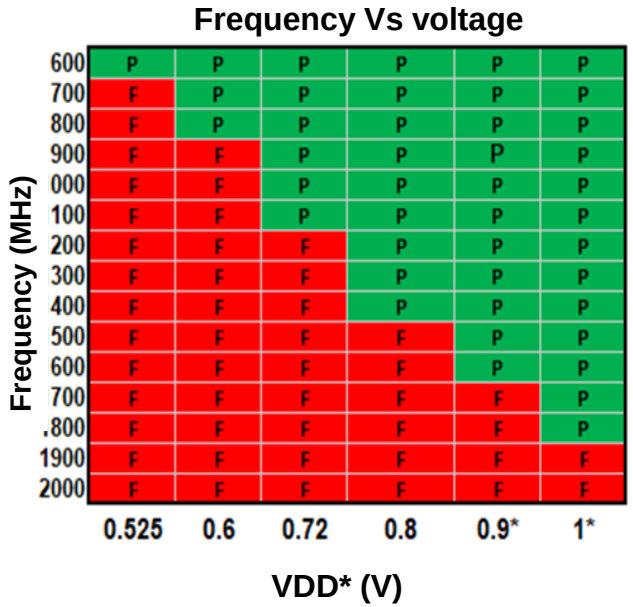


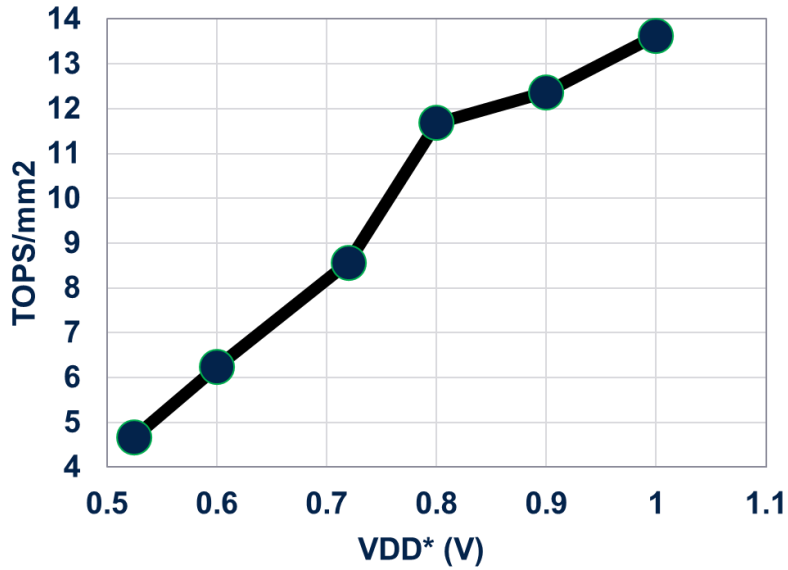
1: (X)S_(Y)T : X is the % sparsity in the Kernel data,Y is the % Inter Kernel transition density

life.augmented

FBB (Forward Body Bias) impact







* 0.9V & 1.0V with 1.2V FBB, others with 1.5V FBB

Compute density improves with FBB at fixed VDD

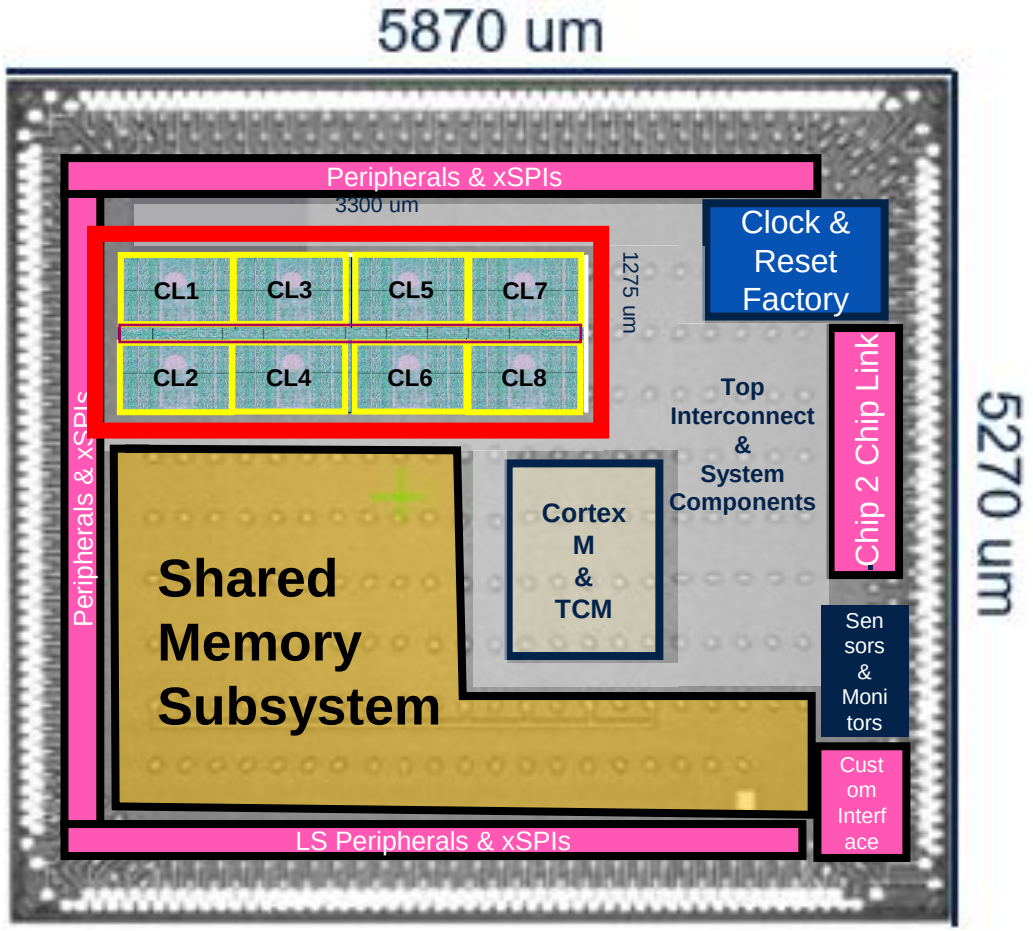
TOPS/mm2 improves 1.8X with FBB across 0V to 1.5V

TOPS/W degrades by only 14% across 1.5V FBB range

life.augmented

Chip summary

Technology 18nm FDSOI
Multi-Cluster IMNPU along with system interconnect: 4.2 mm^2
Voltage range: 0.525-1.0V, FBB 0-1.5V
IMC Capacity 2 Mb
Computation: Deterministic
Precision Mode: 1-4 bits
229 TOPS (Peak Performance) 1 bit Weight - 1bit Feature 57 TOPS (Peak Performance) 4bit Weight - 4bit Feature
310 TOPS/W (1 bit) 77 TOPS/W (4 bit)
54 TOPS/mm2 (1 bit) 13.6 TOPS/mm2 (4 bit)
CNN, LSTM, RNN



CL: IMNPU Cluster

Outline

- Introduction
- In Memory NPU architecture
- SRAM DIMC tile
- Silicon results
- **Mapping strategies**
- Inference examples
- Conclusions

life.augmented

Mapping strategies and optimization for IMC

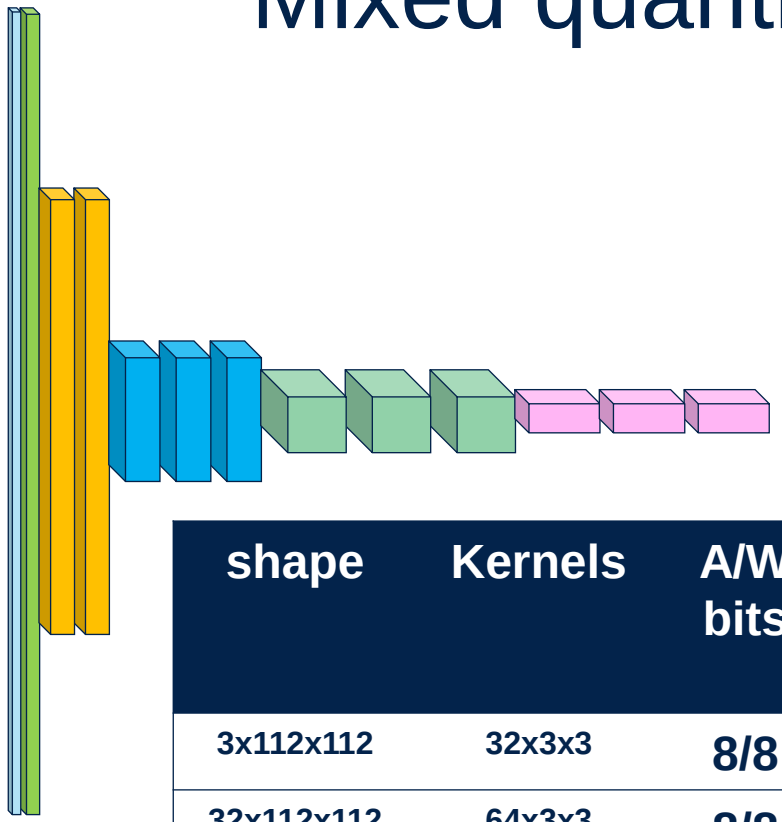
- Quantization 1-8 bits (fixed point today, possibly block scaling next)
- Weight compression/on-the-fly decompression
- Feature-maps compression/decompression
- Structured sparsity
- Layer fusion
- Layer slicing and partitioning: increase parallelism, bandwidth/memory footprint reduction
  - Kernelwise
  - Depthwise
  - Striping
  - Striding

  } **Can be combined**
- Kernel and feature broadcasting, layout optimization, and reloads reduction

IMC can deliver massive amounts of OPS/cycle if data movement is reduced
→major bottleneck

life.augmented

Mixed quantization precision example: VGG16 tiny



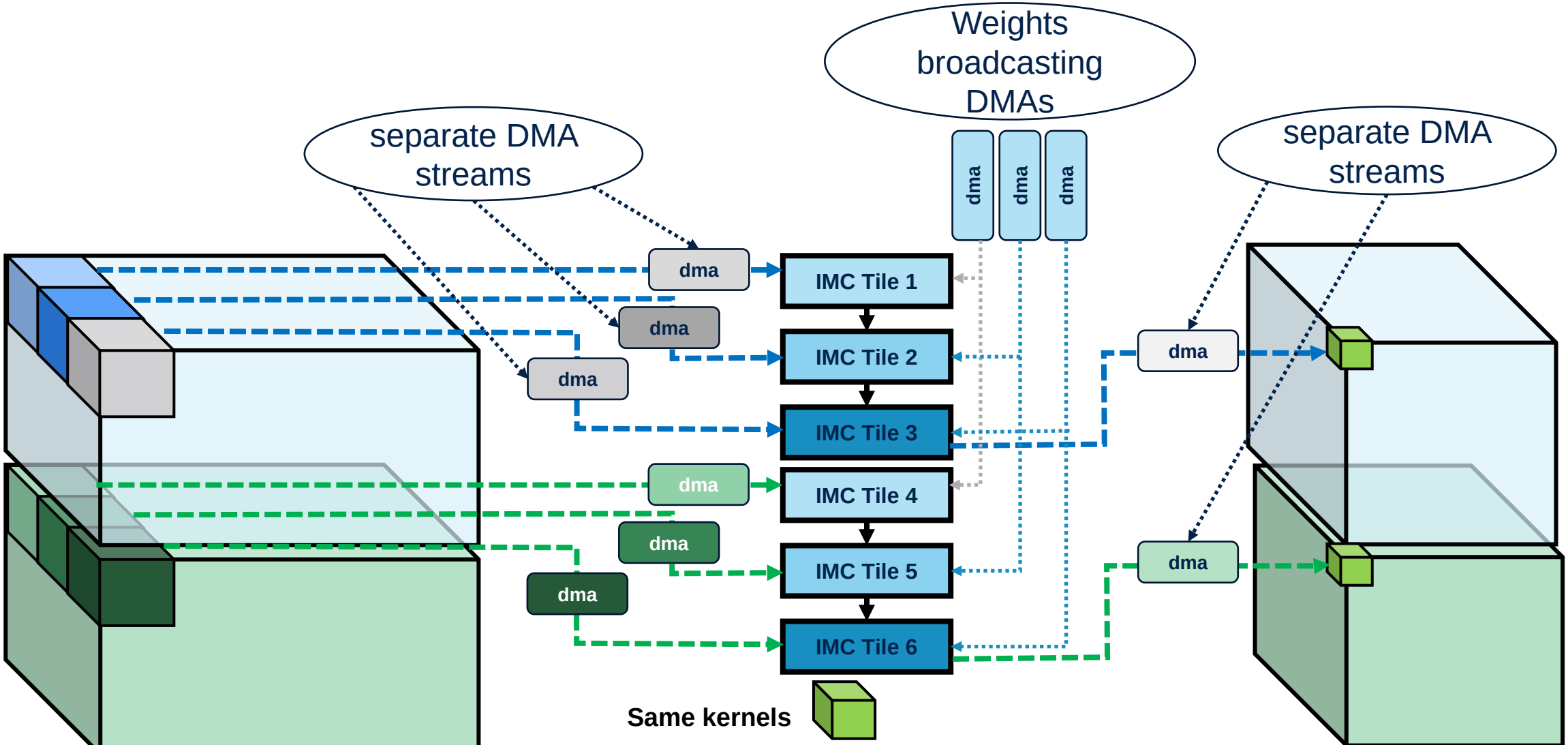
shape	Kernels	A/W bits		A/W bits	Feat comp. (*)	Weights comp.	Total comp. perf
3x112x112	32x3x3	8/8	Mixed Precision Mapping →	8/8	1x	1x	~7x 2.1x
32x112x112	64x3x3	8/8		4/4	2x	2x	
64x56x56	112x3x3	8/8		4/4	2x	2x	
112x28x28	224x3x3	8/8		2/2	4x	4x	
224x7x7	224x3x3	8/8		1/1	8x	8x	

Accuracy loss 1-3% →

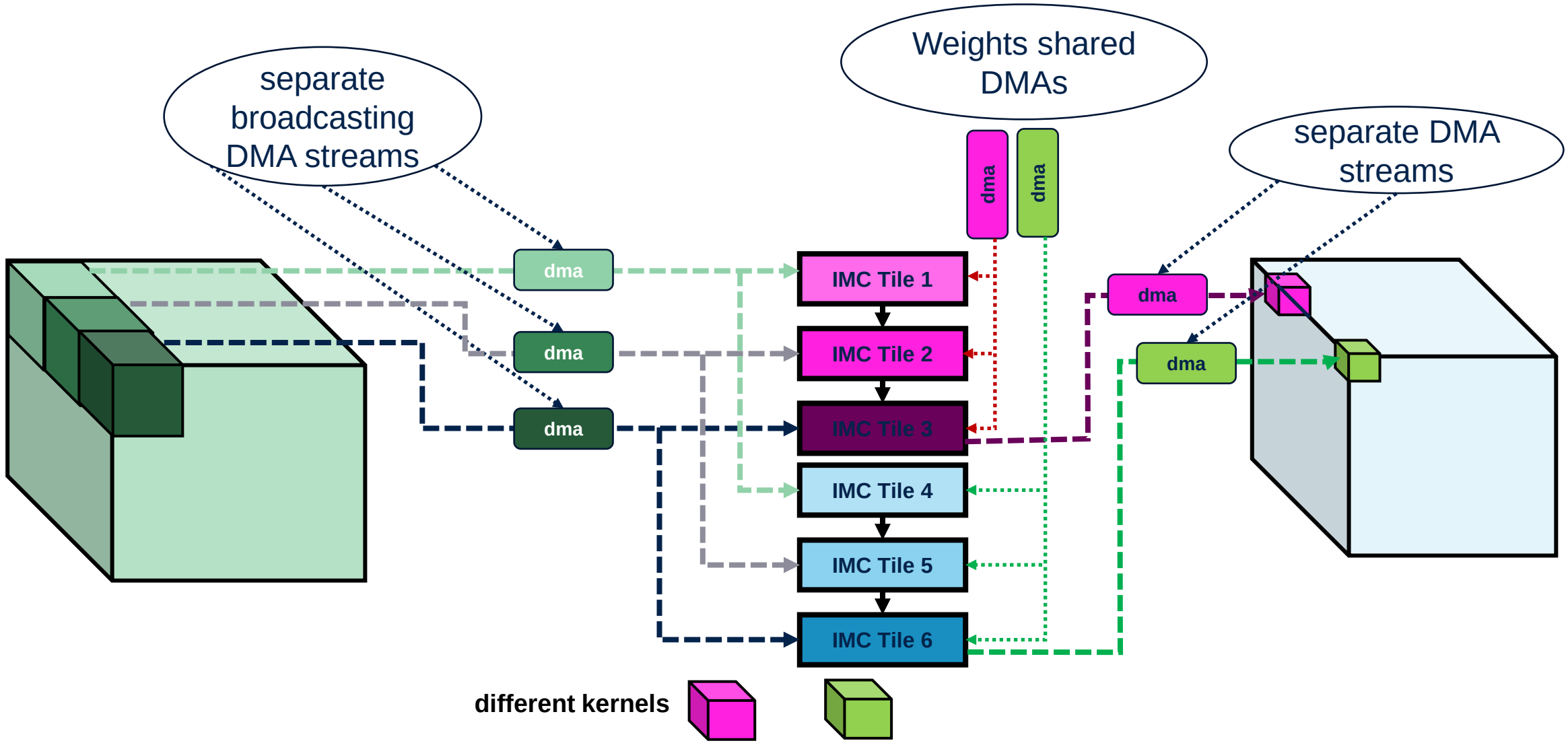
(*) actual feature map compression and throughput reduction depend on mapping

life.augmented

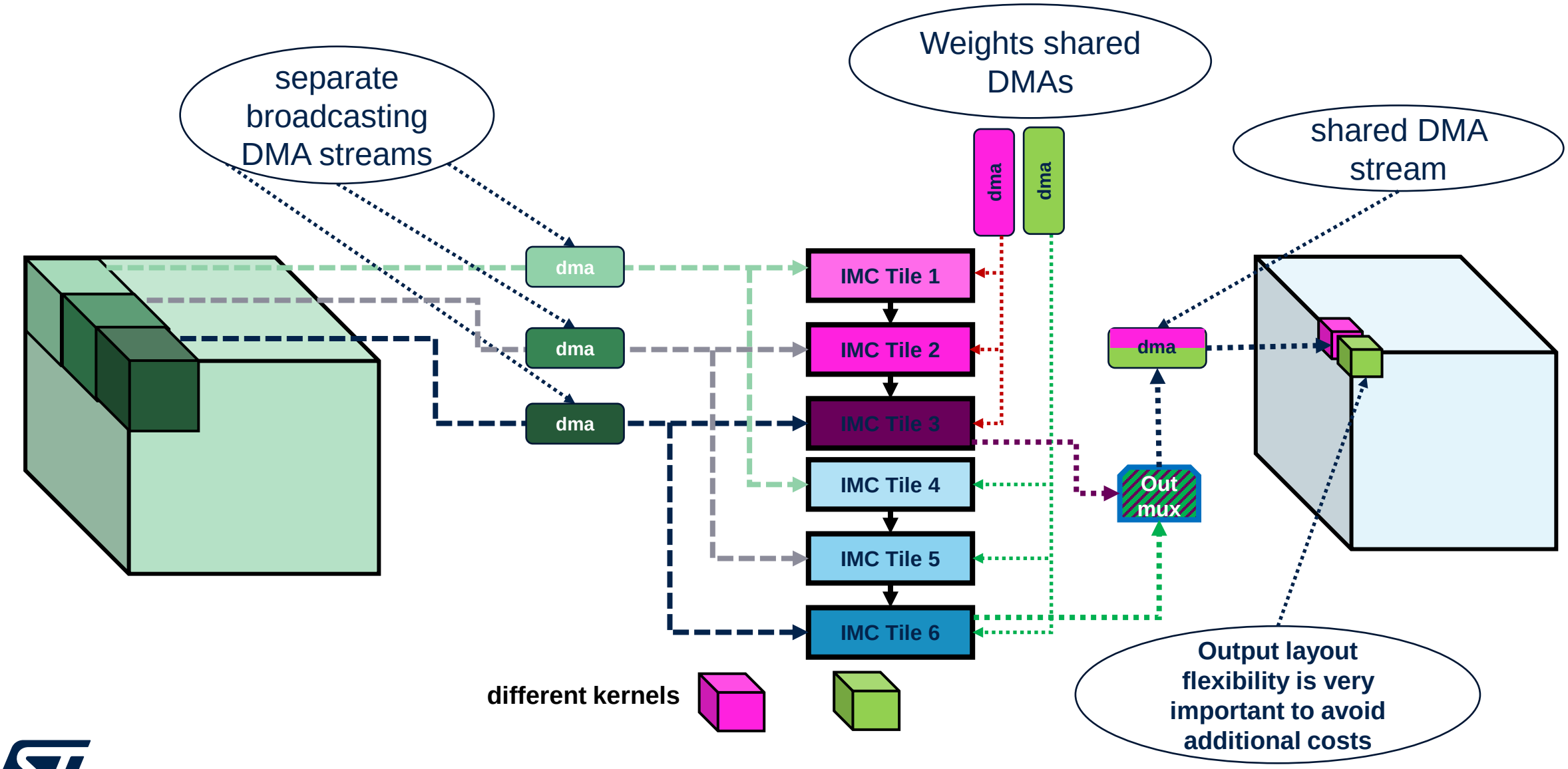
Layer partitioning: chaining & striping



Layer partitioning: chaining & kernelwise



Layer partitioning: chaining & kernelwise



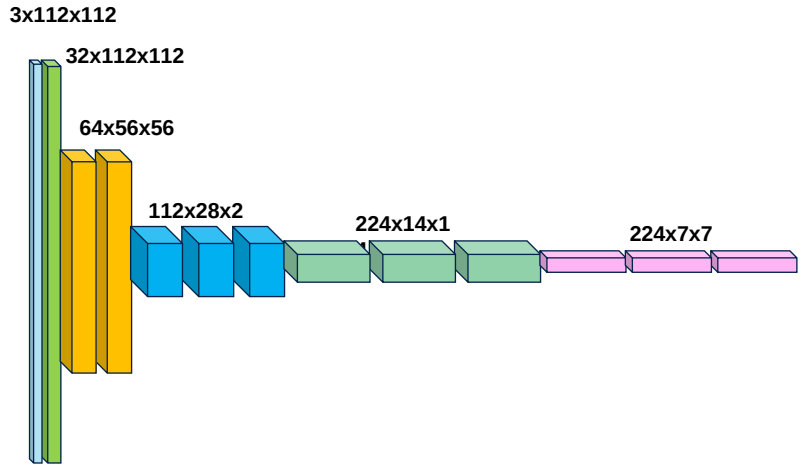
Outline

- Introduction
- In Memory NPU architecture
- SRAM DIMC tile
- Silicon results
- Mapping strategies
- **Inference examples**
- Conclusions

life.augmented

VGG16 style network mapping

layer	CxHxW	no of kernels	no of params	activation (bytes)	weigths (bytes)	kernel (bits)	MMACs	stripes chains parallel	IMC utilization	kernel rounds	cycles	
									1 cluster			8 clusters[1]
C1_1	3x112x112	32	896	18816	432	108	10.84	8,1,1	11%	1	5346	2352
C1_2	32x112x112	64	18496	200704	9216	1152	231.21	4,2,1	56%	2	114048	14688
C2_1	64x56x56	64	36928	100352	18432	2304	115.61	2,3,1	56%	2	77568	10560
C2_2	64x56x56	112	64624	100352	32256	2304	202.31	1,3,2	56%	2	135744	18480
C3_1	112x28x28	112	113008	43904	56448	4032	88.51	1,4,2	98%	2	50274	8930
C3_2	112x28x28	112	113008	43904	56448	4032	88.51	1,4,2	98%	2	50274	8930
C3_3	112x28x28	224	226016	43904	112896	4032	177.02	1,4,2	98%	4	100548	14333
C4_1	224x14x14	224	451808	21952	225792	8064	88.51	1,8,1	98%	7	71442	10206
C4_2	224x14x14	224	451808	21952	225792	8064	88.51	1,8,1	98%	7	71442	10206
C4_3	224x14x14	224	451808	21952	225792	8064	88.51	1,8,1	98%	7	71442	10206
C5_1	224x7x7	224	451808	5488	225792	8064	22.13	1,8,1	98%	7	39029	5576
C5_2	224x7x7	224	451808	5488	225792	8064	22.13	1,8,1	98%	7	39029	5576
C5_3	224x7x7	224	451808	5488	225792	8064	22.13	1,8,1	98%	7	39029	5576



Configuration	1 cluster	8 clusters
MACS/inf	1.25E+09	
cycles/inf	865214	125618
Inf/sec	693	4776
TOPS/W[2]	46.8	

Measured at 0.525V and 600MHz with 1.5v FBB

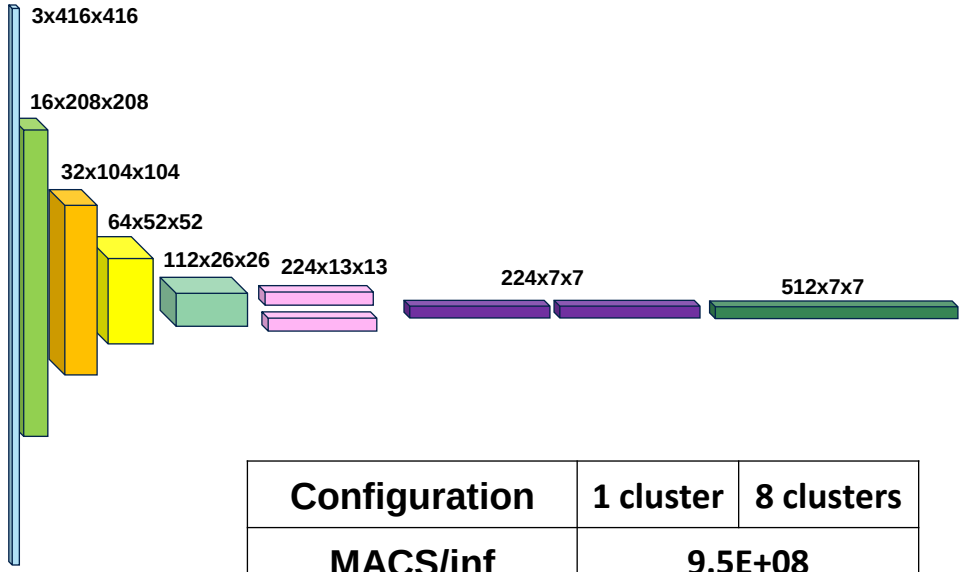
(1) Estimated assuming additional striping and kernels broadcasting, kernel load cycles included

(2) kernels randomly chosen with 50% sparsity

life.augmented

YOLO2 tiny style mapping example

layer	CxHxW	no of kernels	no of params	activation (bytes)	weigths (bytes)	kernel (bits)	MMACs	stripes chains parallel	IMC utilization	kernel rounds	cycles	
									1 cluster			8 clusters[1]
C1	3x416x416	16	448	259584	216	108	74.76	8,1,1	5%	1	36531	4590
C2	16x208x208	32	4640	346112	2304	576	199.36	8,1,1	56%	1	97632	12456
C3	32x104x104	64	18496	173056	9216	1152	199.36	2,2,2	56%	1	98496	13320
C4	64x52x52	112	64624	86528	32256	2304	174.44	1,3,2	56%	2	117600	16212
C5	112x26x26	224	226016	37856	112896	4032	152.64	1,4,2	98%	4	88641	12844
C6_1	224x13x13	112	225904	18928	112896	8064	38.16	1,8,1	98%	4	32744	5857
C6_2	224x13x13	112	225904	18928	112896	8064	38.16	1,8,1	98%	4	32744	5857
C7	224x7x7	224	451808	5488	112896	4032	22.13	1,4,2	98%	4	19514	4879
C8	224x7x7	512	1E+06	5488	258048	4032	50.58	1,4,2	98%	8	44604	5576
C9	512x7x7	30	15390	12544	69120	2048	0.75	1,2,4	100%	1	9008	9008



Configuration	1 cluster	8 clusters
MACS/inf	9.5E+08	
cycles/inf	577514	90598
Inf/sec	1039	6623
TOPS/W[2]	50.86	

Measured at 0.525V and 600MHz with 1.5v FBB

(1) Estimated assuming additional striping and kernels broadcasting, kernel load cycles included

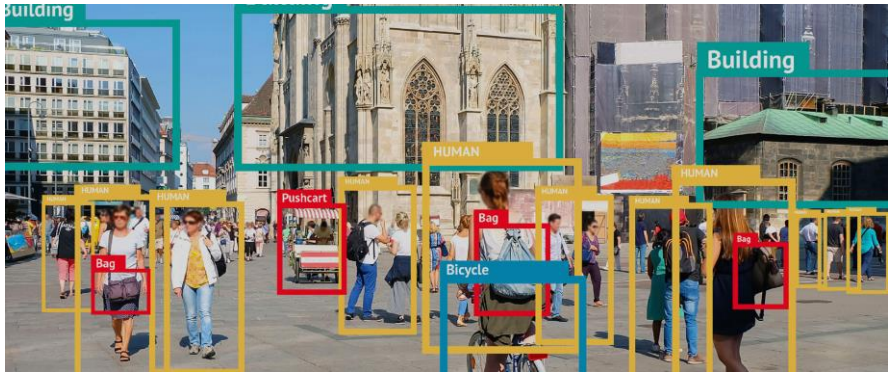
(2) kernels randomly chosen with 50% sparsity

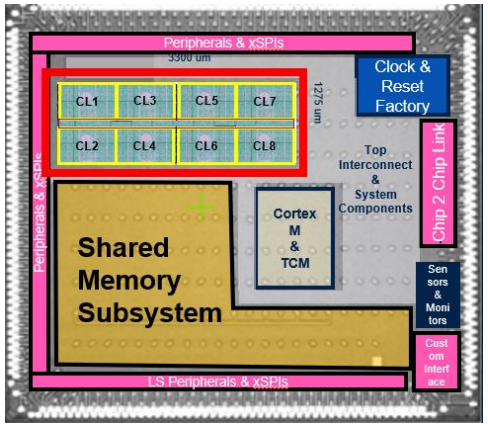
life.augmented

Battery-operated device for video surveillance

Configuration	MACS/ Inference	Inf/s	IMNPU Power	Total Power[1]	Battery endurance[2] (1/100 duty cycle)
1 cluster @ 10 MHz, 0.0V FBB Always-ON,VGG like	1.25 GOPS	10	267 uW	567 uW	363 days
8 clusters @ 400MHz, 0.3V FBB Post Wakeup, 10x complexity	12.5 GOPS	30	8.0 mW	12.0 mW	

(1) Estimated power includes a portion of shared memory, IOs, clock, and external sensor interface, weights stored in ePCM on chip
(2) 6000 mA/h battery capacity assumed (e.g., 2 AA 1.5v batteries)





In Memory NPU sub-system example, fixed Vdd, multiple Body Bias island

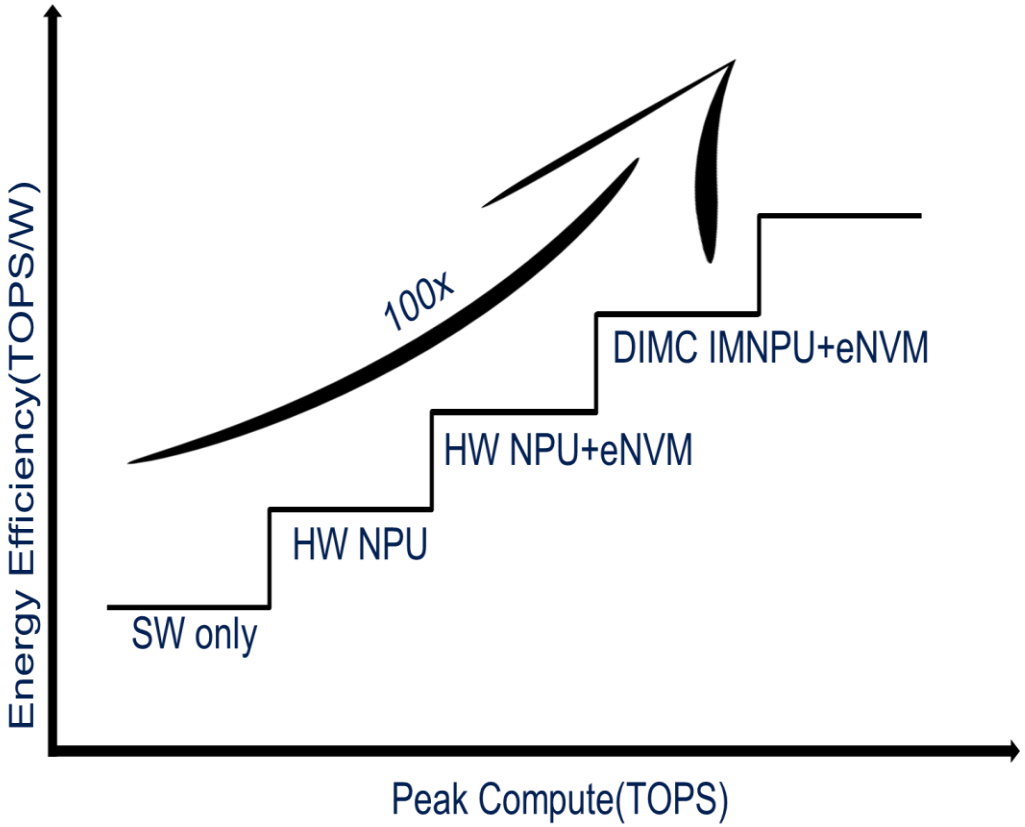
life.augmented

Outline

- Introduction
- In Memory NPU architecture
- SRAM DIMC tile
- Silicon results
- Mapping strategies
- Inference examples
- **Summary**

life.augmented

Conclusions



- Embedded NPUs are enabling efficient NN inference on the edge
- In Memory Computing is a key enabler to achieving higher compute density and energy efficiency: our results in 18nm FD-SOI show up to **50x** improvements compared to pure digital logic
- DIMC-based NPU maintains deterministic computation → general-purpose
- Dedicated compilation and optimization tools are key to efficiently mapping the NN computations on these architectures

life.augmented

Our technology starts with You

Find out more at www.st.com

life.augmented

Copyright Notice

www.tinyml.org

Copyright Notice

www.tinyml.org