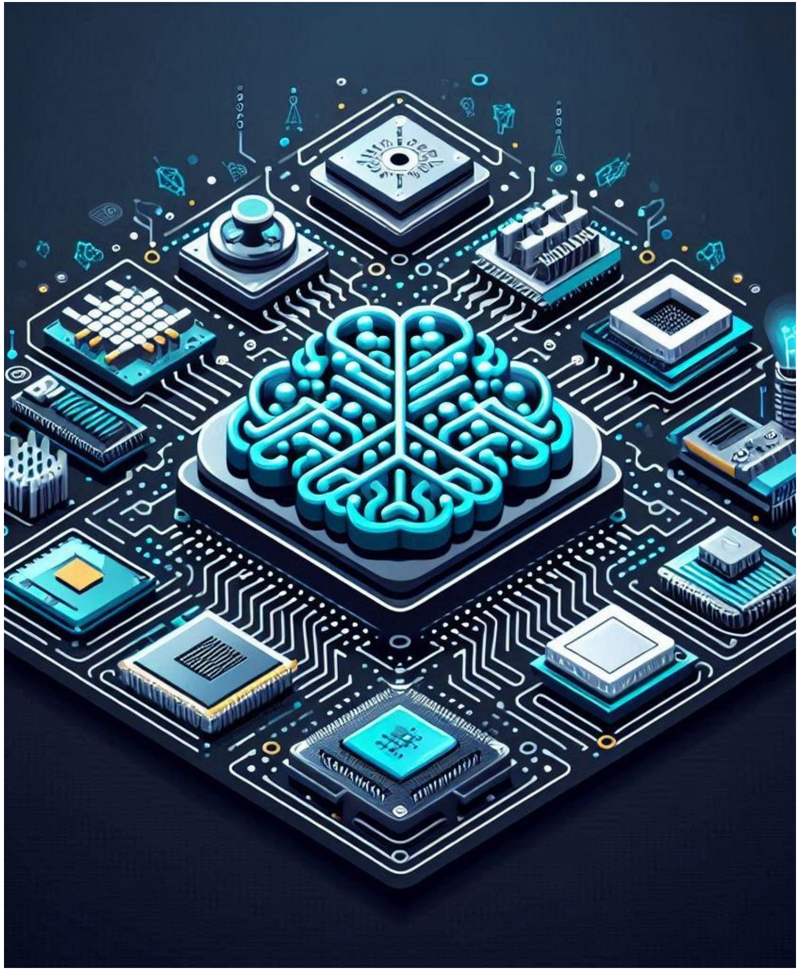


Image generated by Microsoft Copilot.

CHALMERS

NextGen Accelerators: Flexible, Scalable, Efficient – Together^3

Pedro Petersen Moura Trancoso

Full Professor, Computer Science and Engineering

Chalmers University of Technology

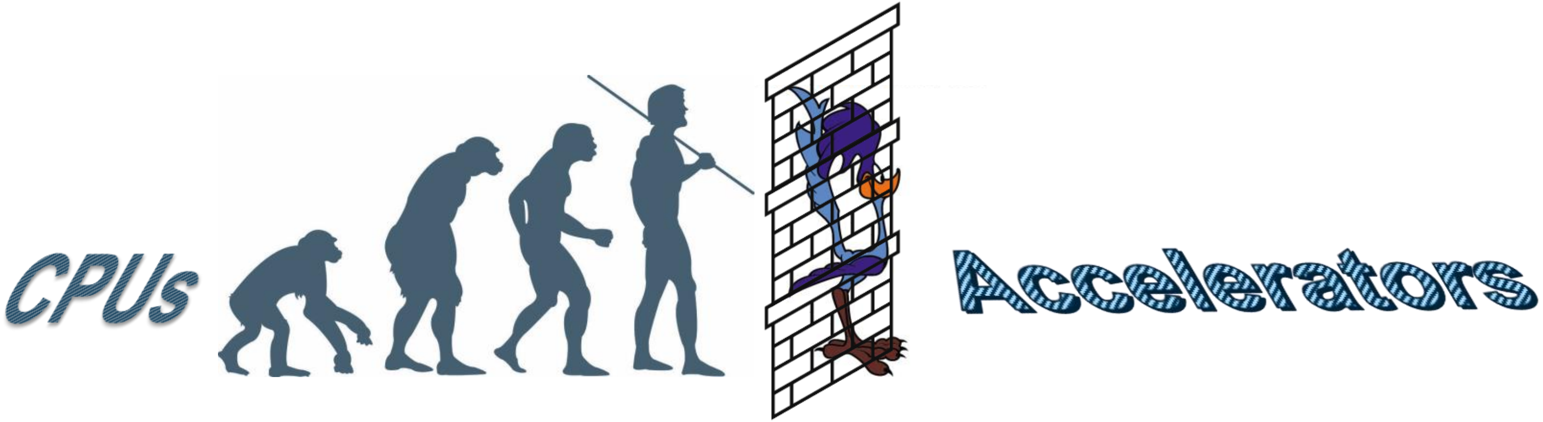
Gothenburg, Sweden

NextGen Accelerators: Flexible, Scalable, Efficient – Together^3

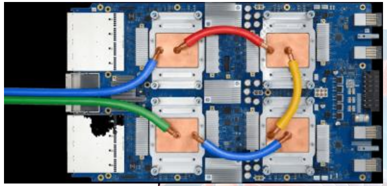
https://www.scientificamerican.com/article/what-causes-the-feeling-of-deja-vu/

All has been said before!

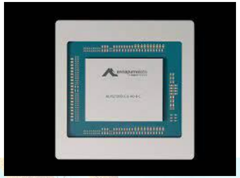
Motivation…



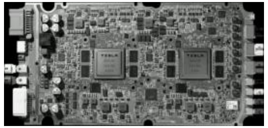
Accelerators



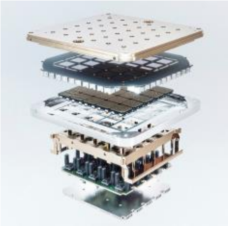
Google TPU



Amazon Inferentia & Tranium



Tesla FSD



Tesla Dojo D1

NVIDIA Jetson Orin

Graphcore IPU

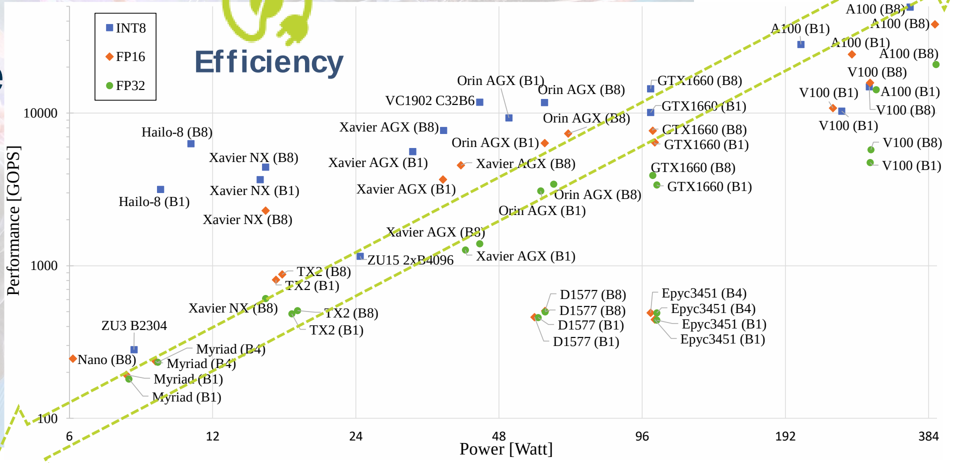
AMD Versal

Hailo-8

High-performance

Efficiency

Low power

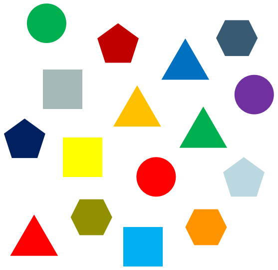


Results from VEDLIoT D3.3

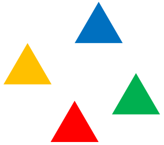
Cambrian explosion: canbedone / Adobe Stock

Accelerators design tradeoff

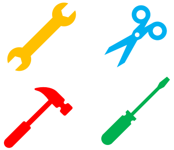
Applications



Select a Domain

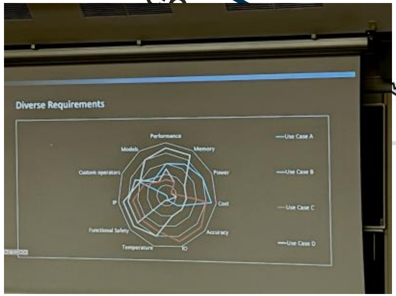


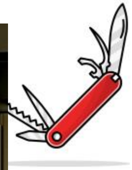
Dedicated



- ☑ High efficiency
- ☒ Flexible

Gen



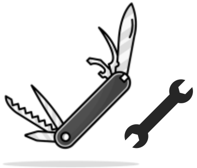


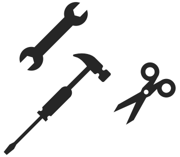
- ☒ High efficiency
- ☑ Flexible

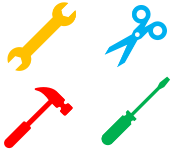
I WANT BOTH!!!

Swiss Army Knife and Crying Boy images by Vectorportal.com

Accelerators design







From generic to dedicated

From dedicated to generic

Automatic Binding Blocks, 1949

Accelerators design how-to…

https://twitter.com/IKEAUK/status/1252269467515617280?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1252269467515617280%7Ctwgr%5E%7Ctwcon%5Es1_c10

Our work (so far…)
Deep Learning accelerators

Design → Build → Run

Our work (so far…)
Deep Learning accelerators

Design → Build → Run

- High-level estimation co-design tool - RAINBOW
- Roofline-based framework

Our work (so far…)

Deep Learning accelerators

Design → Build → Run

- Resource management – ARADA
- Scratchpad Memory Management

Our work (so far…)

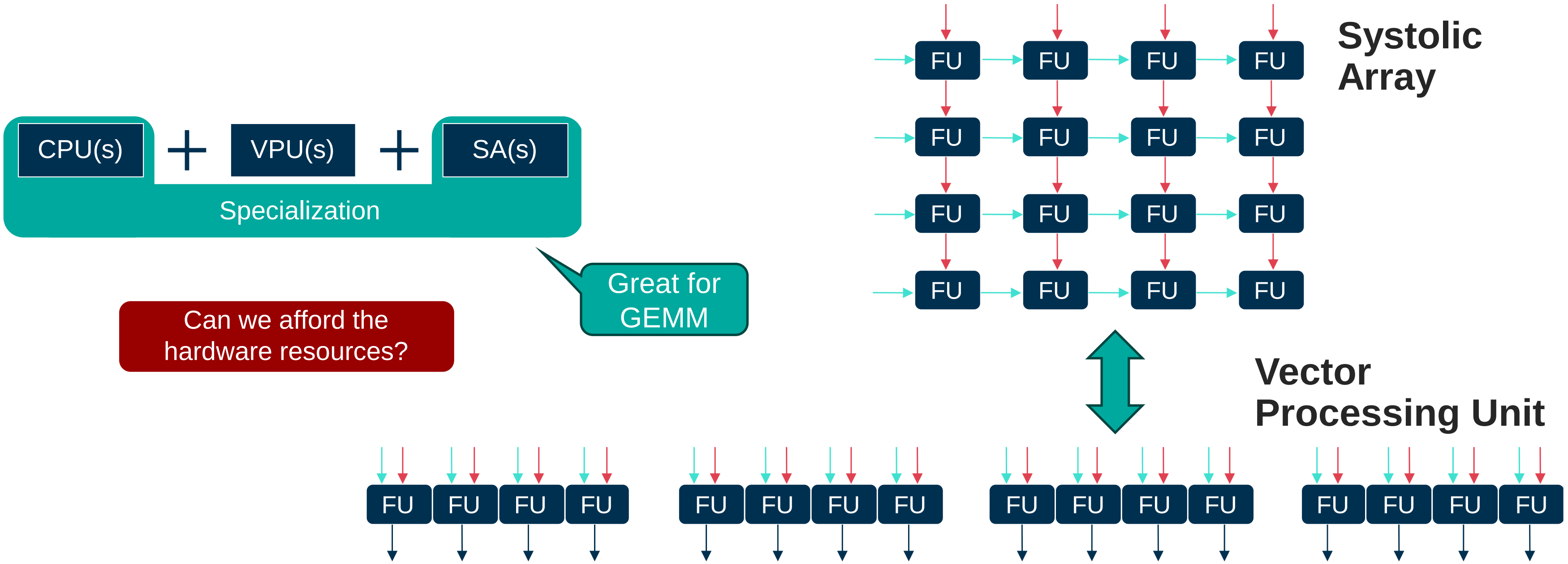
Deep Learning accelerators

Design → Build → Run

- From generic to dedicated – VSA
- From dedicated to generic – FiBHA

From generic to dedicated

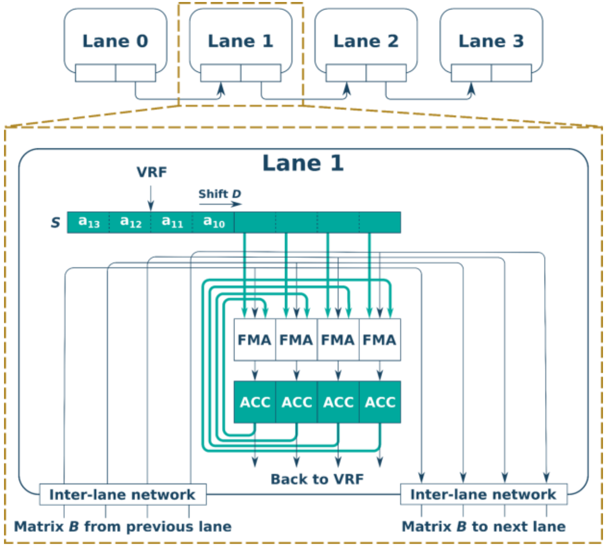
VSA: A Hybrid Vector-Systolic Architecture



- M. V. Maceiras, M. Waqar Azhar and P. Trancoso, "VSA: A Hybrid Vector-Systolic Architecture," *2022 IEEE 40th International Conference on Computer Design (ICCD)*, Olympic Valley, CA, USA, 2022, pp. 368-376
- M. V. Maceiras, M. W. Azhar and P. Trancoso, “Exploiting the Potential of Flexible Processing Units.” In Proc. of the IEEE 35th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD 2023), pp. 1-12

From generic to dedicated

VSA hardware, software, experimental setup



RISC-V®

Algorithm 2 GEMM using custom instruction

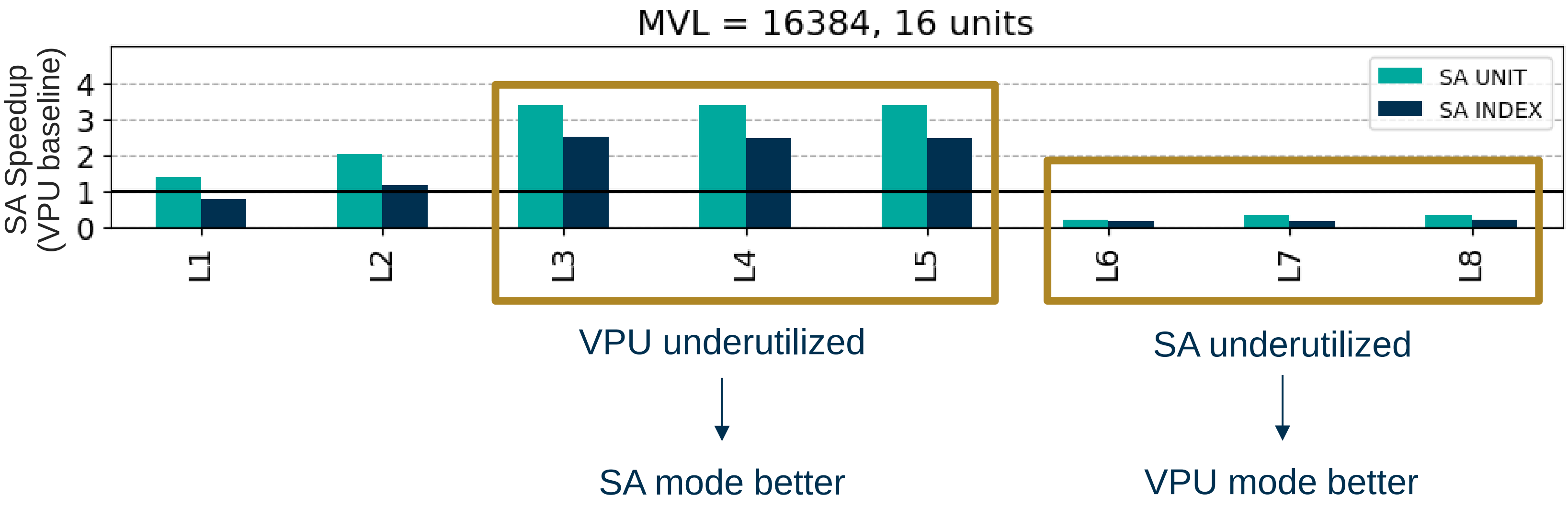
```
for all i ∈ {1, ..., M/SA_R} do
    v_r = LOAD_ROW_SET(i)
    for all j ∈ {1, ..., N/SA_C} do
        v_c = LOAD_COL_SET(i)
        v_t = INIT_TILE(i,j)
        v_t = SA(v_r, v_c, v_t)
    end for
    STORE(v_t)
end for
```

"Magic sauce" – hardware overhead < 0.1% area

Experimental Setup:

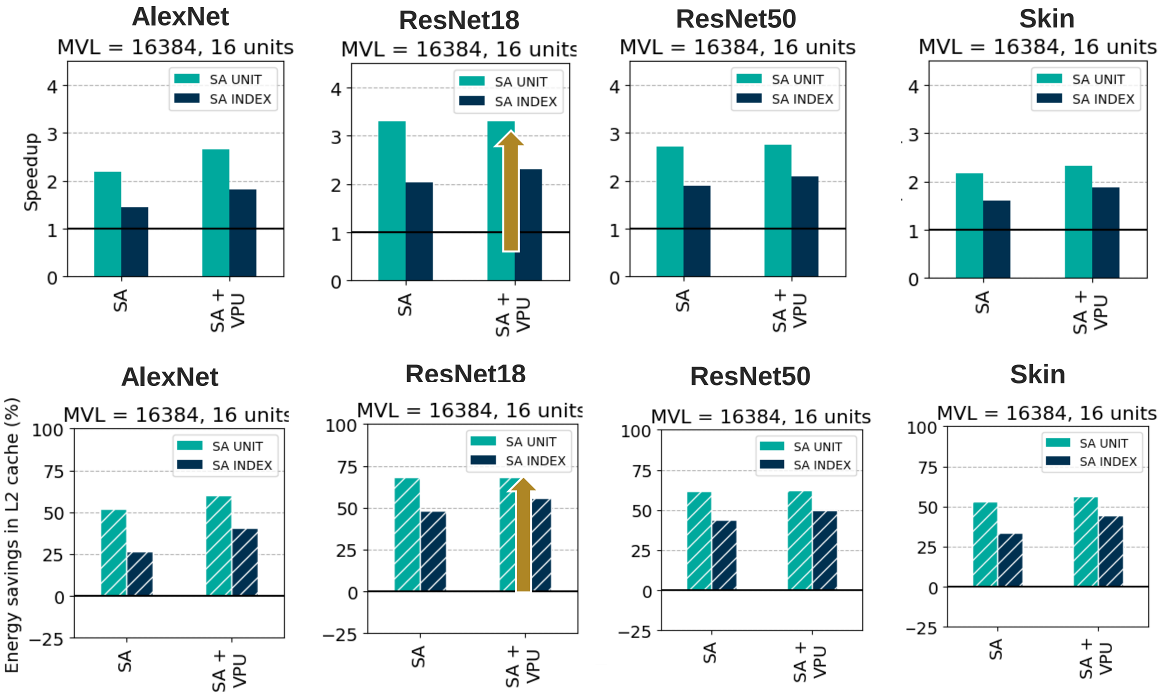
- RISC-V VPU
- Simulation: gem5+McPAT
- Implementation (eProcessor / 65nm)
- Index and unitary data load
- Workloads: AlexNet, ResNet18/50, Skin (DeepHealth)

From generic to dedicated
Vector Processing Unit vs. Systolic Array



Different phases of same application benefit differently from VPU or SA – Hybrid can achieve the best of both worlds!

From generic to dedicated
VSA speedup and energy savings



Minimal area overhead of 0.1%

Up to 3.5x speedup

Up to 70% energy savings in cache

From generic to dedicated

Open questions…

Which are quality metrics? (performance/area)

Which extensions make sense?

???

How should we configura a multi-engine accelerator?

How many extensions to be supported at the same time?

How dedicated should a generic engine be?

Our work (so far…)
Deep Learning accelerators

Design → Build → Run

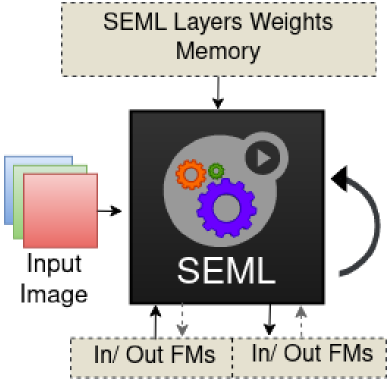
- From generic to dedicated – VSA
- From dedicated to generic – FiBHA

From dedicated to generic

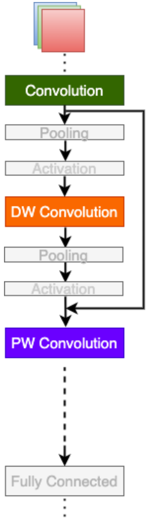
FiBHA: Fixed Budget Hybrid CNN Accelerator

Observation: ML Models are increasingly and more heterogenous

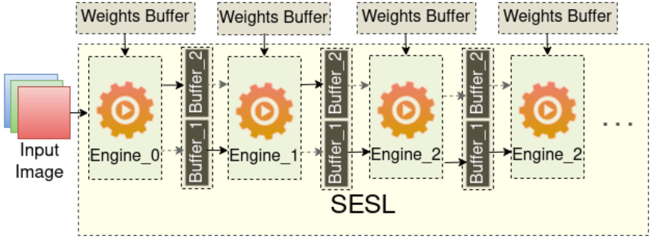
Option A



All layers are executed by the same engine
Single Engine Multiple Layers (SEML)



Option B



Each layer is executed by a different engine
Single Engine Single Layer (SESL)

- F. Qararyah, M. W. Azhar and P. Trancoso, "FiBHA: Fixed Budget Hybrid CNN Accelerator," *2022 IEEE 34th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, Bordeaux, France, 2022, pp. 180-190
- F. Qararyah, M. W. Azhar, and P. Trancoso, “An Efficient Hybrid Deep Learning Accelerator for Compact and Heterogeneous CNNs,” ACM Transactions on Architecture and Code Optimimization (TACO) 21(2), Article 25 (June 2024), 26 pages.

From dedicated to generic

SESL & SEML: When to use which?

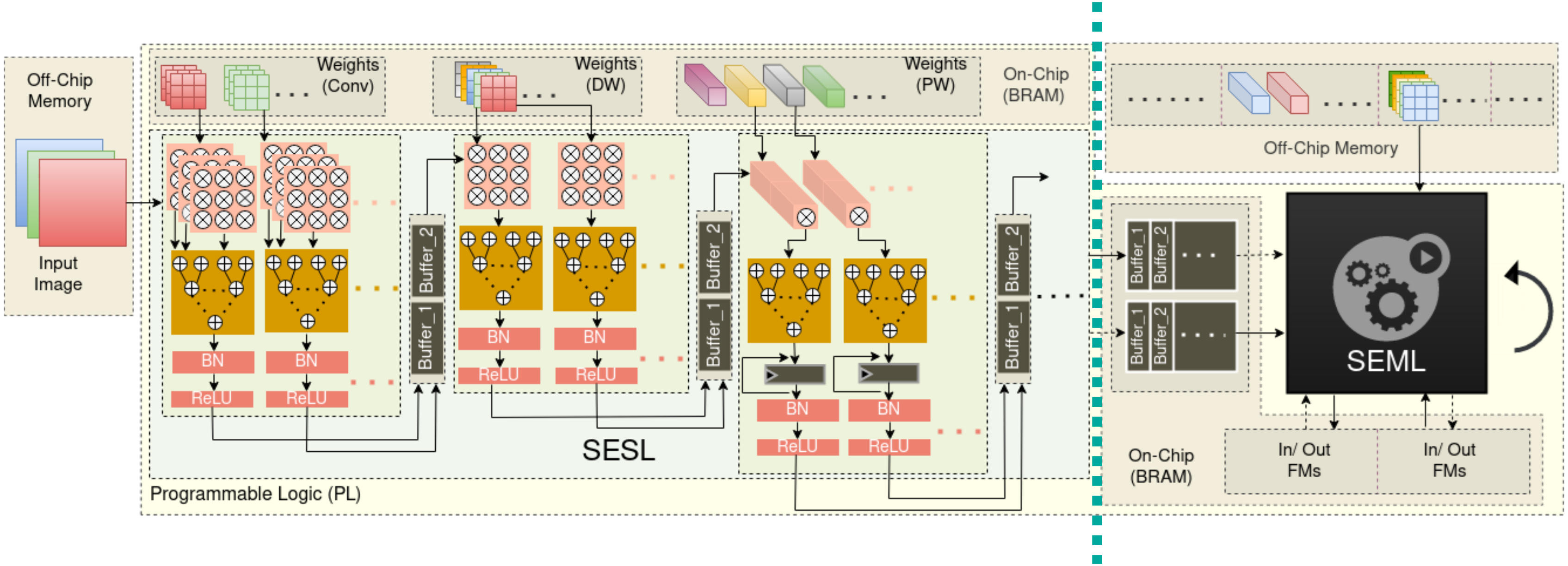
Keep weights in local memory

SplitCNN

Limited local memory space

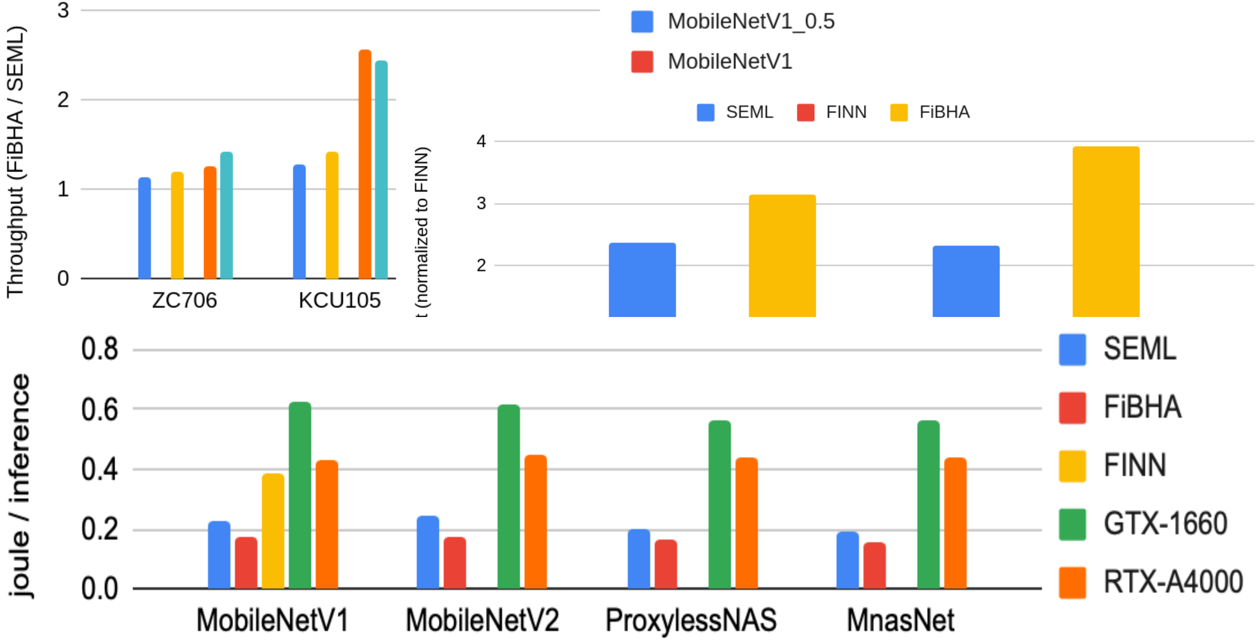
Dataflow reduce temporary storage

From dedicated to generic
FiBHA example



Implemented in HLS, evaluated on FPGA!

From dedicated to generic
Results



FiBHA hybrid accelerator balances heterogeneity & resource budget

≅**4x** Throughput improvement

≅**2x** Energy efficiency

From dedicated to generic

Open questions…

Which combinations of engines and configurations into a multi-engine accelerator?

Which engines should be made available?

???

Which configurations depending on goals?

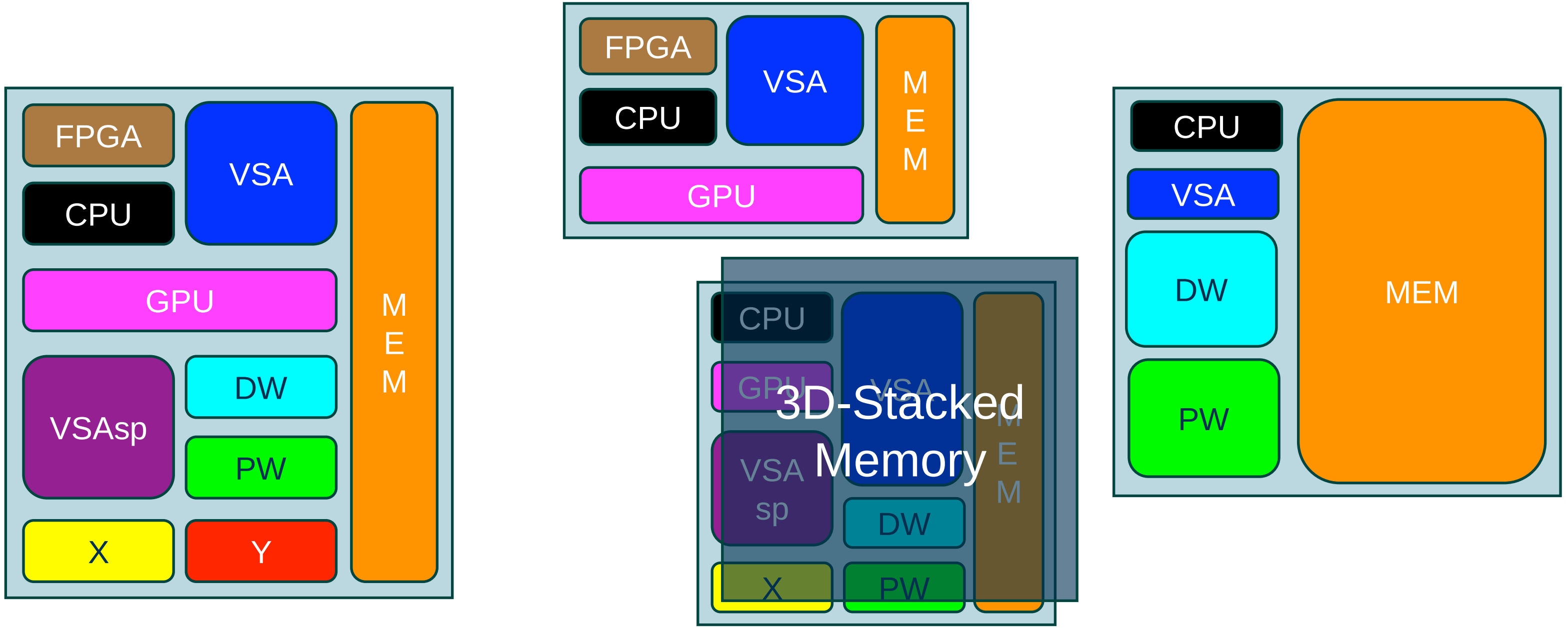
How generic should a dedicated engine be?

“Science Fiction” => The Vision

Flexible, Scalable, Efficient – Together[1-2-3]

Volvo Museum, Gothenburg, Sweden

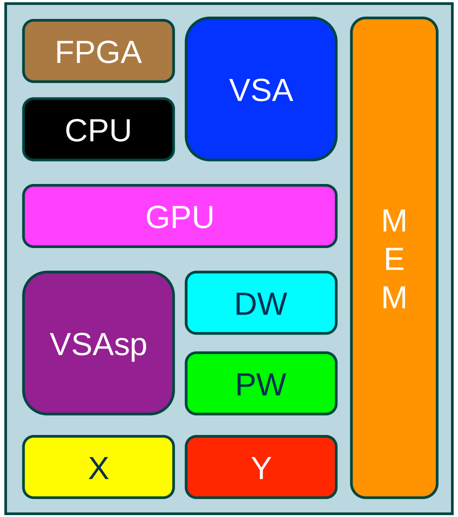
The Vision
Flexible, Scalable, Efficient – Together[1]

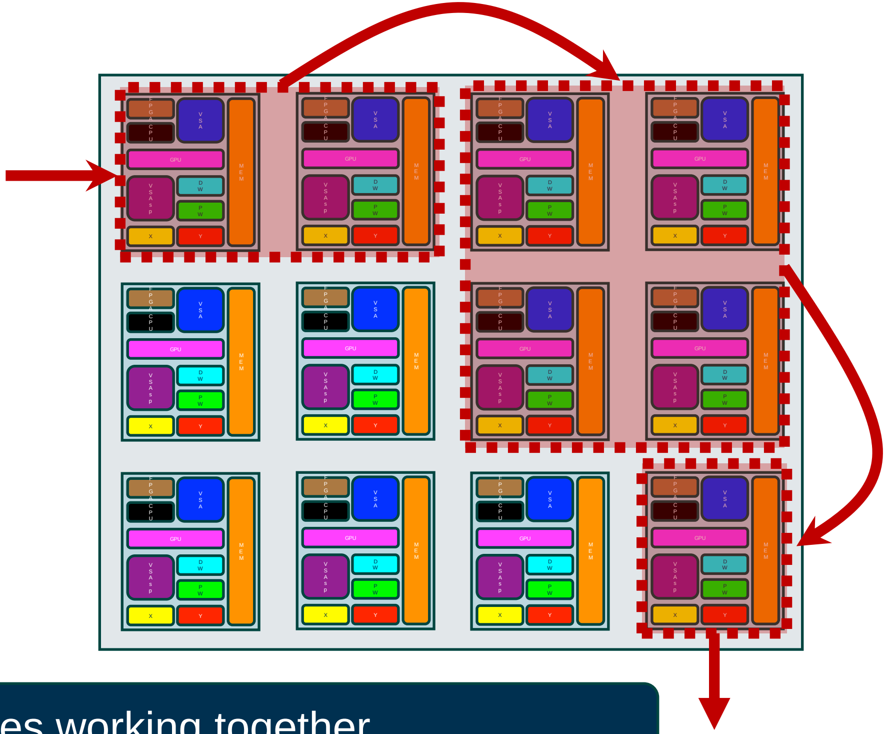


Units working together

The Vision

Flexible, Scalable, Efficient – Together²

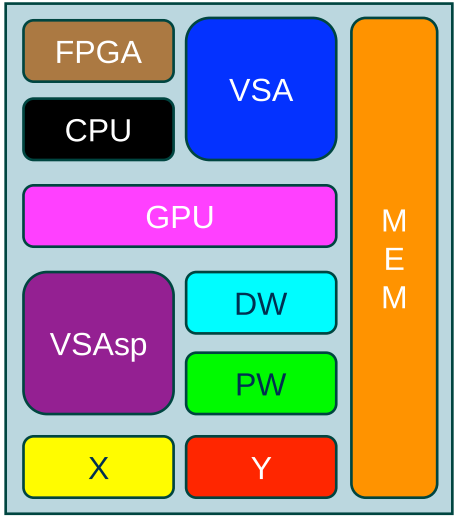


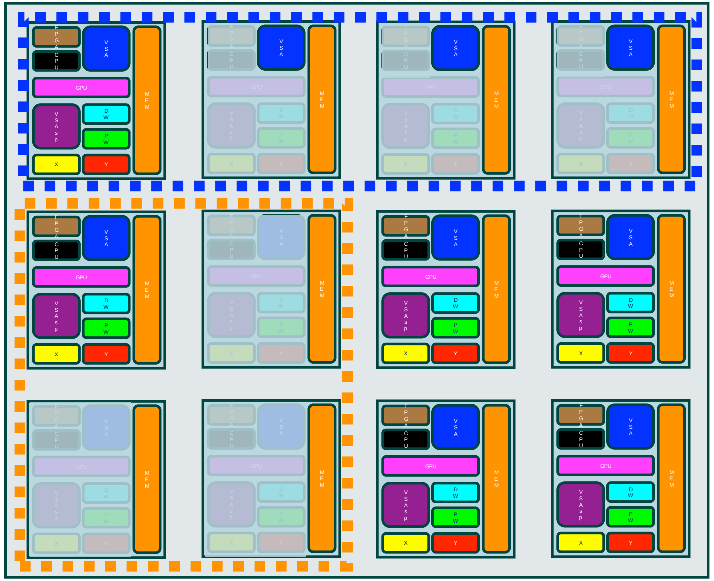


Tiles working together

The Vision

Flexible, Scalable, Efficient – Together³



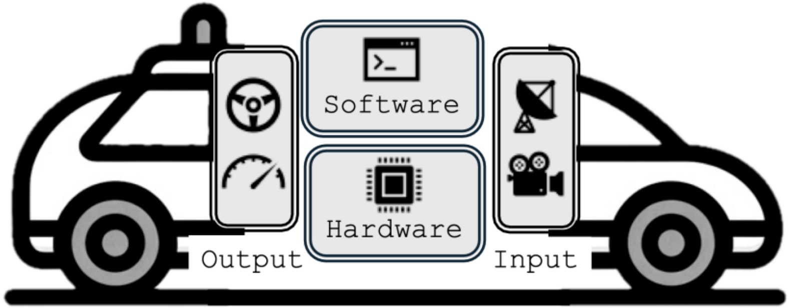


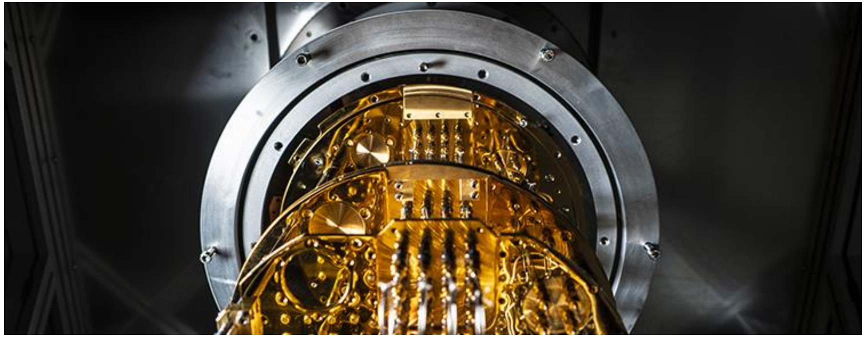
Resources shared together

We also accelerate…

Swedish Foundation
for Strategic Research

CHALMERS





SSF-AutoPiM

- Develop energy-efficient hardware accelerators for autonomous vehicles
- Deep learning application
- Novelty: combine near- and in-memory proc.
- Our contribution: near-memory processing
- Collaboration: Bar-Ilan University

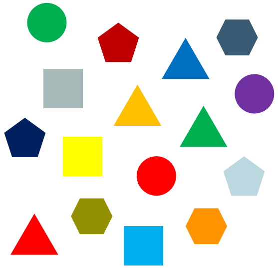
X. Wang, M. A. Maleki, M. W. Azhar, P. Stenström, and P. Trancoso, "Challenges and Directions for Autonomous Driving Hardware Accelerators", ACACES 2024, Italy, July 2024

SSF-QuantumStack

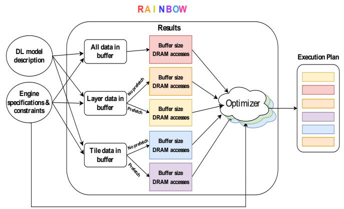
- Develop a full software stack for programming quantum computers
- Novelty: Bring all together – physics, computer science and engineering; improve programmability for QC
- Our contribution: hardware acceleration for QC simulation and error correction
- Collaboration: CSE and WACQT

Conclusions

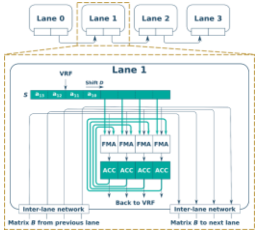
Applications



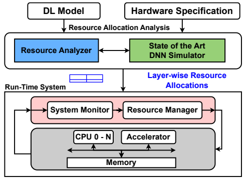
RAINBOW

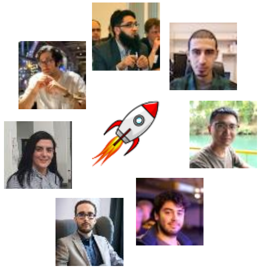


VSA

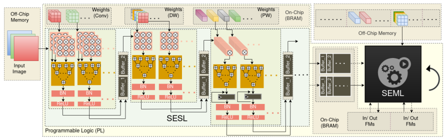


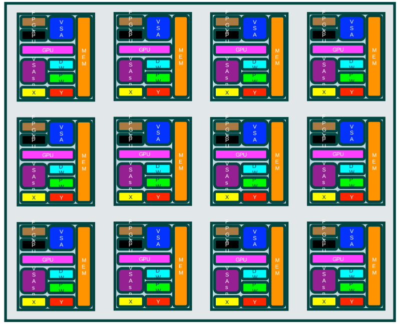
ARADA





FiBHA





VEDLIoT
Very Efficient Deep Learning in IoT

eProcessor

European Processor Initiative

EuroHPC
Joint Undertaking

European Commission

Flexible, Scalable, Efficient – Together³

AVANCEZ

1829

CHALMERS